Structured Sparsity via Optimal Interpolation Norms

Andrew Michael McDonald

First Supervisor: Massimiliano Pontil
Second Supervisor: Mark Herbster

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Computer Science
University College London

I, Andrew Michael McDonald, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

We study norms that can be used as penalties in machine learning problems. In particular, we consider norms that are defined by an optimal interpolation problem and whose additional structure can be used to encourage specific characteristics, such as sparsity, in the solution to a learning problem.

We first study a norm that is defined as an infimum of quadratics parameterized over a convex set. We show that this formulation includes the k-support norm for sparse vector learning, and its Moreau envelope, the box-norm. These extend naturally to spectral regularizers for matrices, and we introduce the spectral k-support norm and spectral box-norm. We study their properties and we apply the penalties to low rank matrix and multitask learning problems.

We next introduce two generalizations of the k-support norm. The first of these is the (k,p)-support norm. In the matrix setting, the additional parameter p allows us to better learn the curvature of the spectrum of the underlying solution. A second application is to multilinear algebra. By considering the rank of its matricizations, we obtain a k-support norm that can be applied to learn a low rank tensor. For each of these norms we provide an optimization method to solve the underlying learning problem, and we present numerical experiments.

Finally, we present a general framework for optimal interpolation norms. We focus on a specific formulation that involves an infimal convolution coupled with a linear operator, and which captures several of the penalties discussed in this thesis. Finally we introduce an algorithm to solve regularization problems with norms of this type, and we provide numerical experiments to illustrate the method.

In memory of my brother James Alexander McDonald.

Acknowledgements

First, I would like to thank my supervisor, Massimiliano Pontil, for his guidance and support throughout the PhD. He has always been very generous with his time, has been a great mentor and I am greatly indebted to him for giving me the chance to learn from him.

I am grateful to my second supervisor, Mark Herbster, with whom I was fortunate to have many interesting research discussions, as well as work on a paper during my time at UCL.

I would like to thank Charles Micchelli and Patrick Combettes, who collaborated with Massi and I on a number of topics in this thesis, as well as Andreas Maurer for interesting discussions.

I would also like to thank current and former students of Massi's with whom I have had the opportunity to discuss research ideas, including Dimitris Stamos, who collaborated on several results in this thesis, Bernardino Romera Parades, Andreas Argyriou, Julien Bohné and Carlo Gilberto.

Finally, I would like to thank my parents and my brother for their love and support.

Contents

List of Figures

List of Tables

List of Symbols

$\mathcal{H}$	a Hilbert space
$\mathcal{X}, \mathcal{Y}$	Euclidean vector spaces
$\mathbb{N}_d$	the set of integers from 1 up to and including d
$\mathbb{R}^d$	the d dimensional real vector space
$\mathbb{R}^d_+$	the set of vectors with nonnegative components
$\mathbb{R}^d_{++}$	the set of vectors with strictly positive components
$\mathbb{R}^{d \times m}$	the space of $d \times m$ real matrices
$\mathbf{S}^d$	the set of real $d \times d$ symmetric matrices
$\mathbf{S}^d_+$	the set of positive semidefinite matrices
$\mathbf{O}^d$	the set of real orthogonal $d \times d$ matrices
$\langle \cdot \mid \cdot \rangle$	the canonical inner product on a Hilbert space
$\lVert \cdot \rVert$	a norm on a Hilbert space
$\lVert \cdot \rVert_*$	the dual norm to $\lVert \cdot \rVert$
$\lVert \cdot \rVert_p$	the ℓ_p-norm, defined for $w \in \mathbb{R}^d$ as $\lVert w \rVert_p = (\sum_{i=1}^d \lvert w_i \rvert^p)^{1/p}$
$\lVert \cdot \rVert_1$	the ℓ_1-norm, defined for $w \in \mathbb{R}^d$ as $\lVert w \rVert_1 = \sum_{i=1}^d \lvert w_i \rvert$
$\lVert \cdot \rVert_2$	the ℓ_2-norm (Euclidean norm), defined for $w \in \mathbb{R}^d$ as $\lVert w \rVert_2 = \sqrt{\sum_{i=1}^d \lvert w_i \rvert^2}$
$\lVert \cdot \rVert_\infty$	the ℓ_p-norm, defined for $w \in \mathbb{R}^d$ as $\lVert w \rVert_\infty = \max_{i=1}^d \lvert w_i \rvert$
Δ^d	the unit d-simplex, defined as $\Delta^d = \{\lambda \in \mathbb{R}^{d+1} \mid \sum_{i=1}^{d+1} \lambda_i = 1\}$
supp$(\cdot)$	the support operator, defined for $w \in \mathbb{R}^d$ as supp$(w) = \{i \mid w_i \neq 0\} \subset \mathbb{N}_d$
card$(\cdot)$	the cardinality operator, defined for $w \in \mathbb{R}^d$ as card$(w) = \lvert$supp$(w)\rvert$
$w_{\mid g}$	the vector $w \in \mathbb{R}^d$ restricted to the support set defined by $g \subset \mathbb{N}_d$
$[w_1, \ldots, w_m]$	the $d \times m$ matrix whose columns are formed by the vectors $w_1, \ldots, w_m \in \mathbb{R}^d$
diag(σ)	the $d \times d$ diagonal matrix having elements σ_i on the diagonal, for a vector $\sigma \in \mathbb{R}^d$
$\succeq$	the positive semidefinite ordering on $\mathbf{S}^d$
tr(W)	the trace of a matrix W
$W^\top$	the transpose of a matrix W
rank$(\cdot)$	the rank operator for matrices
$\sigma(W)$	the vector formed by the singular values of the matrix W

$\lambda(W)$	the vector formed by the eigenvalues of the matrix W
$\|\cdot\|_{\mathrm{fro}}$	the Frobenius norm, defined for $W \in \mathbb{R}^{d\times m}$ as $\|W\|_{\mathrm{fro}} = \sqrt{\sum_{i,j} W_{ij}^2}$
$\|\cdot\|_{\mathrm{tr}}$	the trace norm, defined as the sum of the singular values of a matrix
$\mathrm{conv}(S)$	the convex hull of a subset S of a vector space

Chapter 1

Introduction

In machine learning and statistics we study the problem of learning a model from data for descriptive or predictive purposes. Large amounts of data are becoming available to researchers in a variety of domains, and new technologies are being applied to an ever increasing diversity of fields. As the science of machine learning continues to mature, the complexity of state of the art models tends to increase. Nevertheless, in most domains, the data or the computational resources required to solve the relevant problems limit the complexity of the model we can hope to efficiently and accurately learn. In order to improve the performance of a learning algorithm within these limitations, a standard approach is to directly or indirectly incorporate assumptions on the underlying model into the learning process. By enforcing these structural assumptions, we aim to improve learning accuracy by restricting the model space being explored.

The principle of parsimony known as Occam's razor states that given competing models, the simplest one should prevail, and this has proven to be a powerful heuristic in many scientific fields. In our context of learning potentially complex models from large amounts of data, this principle naturally leads to sparse models, that is, models that depend only on a small number of parameters. In various problem settings this could, for example, translate to vectors with few non zero elements, matrices with low rank, or other, more complex, characteristics.

A priori knowledge about the underlying domain may further suggest some underlying structure for the sparsity pattern, and this can often be incorporated into the problem with the goal of improving learning. The field of structured sparsity studies learning methods in this setting and forms the backdrop for this thesis. Broadly, research around this topic investigates methods to enforce a given set of statistical or structural constraints, optimization methods to solve the resulting problems, and corresponding algorithmic and statistical guarantees. In this thesis we focus primarily on methodology and optimization, rather than on statistical guarantees related to the methods.

1.1 Problem Setting

Our framework is supervised learning, where we have access to data and observations. The training data are drawn randomly from $\mathcal{X} \times \mathcal{Y}$, independently and identically distributed, and used to learn the model h in some Hilbert space $\mathcal{H}$ which predicts $y = h(x) \in \mathcal{Y}$ given input $x \in \mathcal{X}$. The general problem is

$$\min_{h \in \mathcal{H}} \mathcal{L}(h),$$

where the loss function $\mathcal{L} : \mathcal{H} \to \mathbb{R}$ measures the fit of the model to the data. In regression problems, we learn one or more functions which predict real-valued observations, whereas in classification problems, the model allocates a given input to a class chosen from a discrete finite set. In supervised learning, a standard approach is to frame the learning task as an optimization problem that involves minimizing an objective function over the training data with the goal of the model accurately generalizing to unseen data. Furthermore, when domain specific knowledge about the structure of the underlying model is available, encouraging the learned model to have specific properties, such as low complexity, can lead to improved performance.

In this thesis we focus on methods that encode our assumptions via a functional, typically a norm, which serves as a constraint or a regularizer. Our focus throughout is on the penalties, the properties that they encourage in the underlying solution, and optimization methods to solve the corresponding learning problems. We highlight three frameworks that allow us to impose our structural assumptions on the model in this manner. In each case the structure is encouraged using a convex penalty functional $\Omega : \mathcal{H} \to \mathbb{R}$, which is chosen to have the property that desirable solutions (according to our heuristic) take on small values.

In the regularization approach, the problem takes the form $\min \left\{ \mathcal{L}(h) + \lambda \Omega(h) \;\middle|\; h \in \mathcal{H} \right\}$, where the composite objective function involves the loss $\mathcal{L}$ and the penalty Ω, and λ is a positive parameter that regulates the two terms. The problem, also known as Tikhonov regularization, therefore involves a trade off between fitting the data by minimizing $\mathcal{L}$, and imposing some structural assumptions on the solution by minimizing Ω.

A second method is to directly constrain the solution to the optimization problem by solving $\min \left\{ \mathcal{L}(h) \;\middle|\; h \in \mathcal{H}, \Omega(h) \leq \alpha \right\}$, for some positive real α. This constrained formulation, also known as Ivanov regularization, can be interpreted as finding the best fitting model in a restricted solution space.

In the third approach, the optimization problem becomes $\min \left\{ \Omega(h) \;\middle|\; h \in \mathcal{H}, \mathcal{L}(h) \leq \beta \right\}$, for some positive real β. This method, also known as Morozov regularization, can be interpreted as minimizing the complexity of the model subject to a minimum fit to the data.

Originally studied in the inverse problems literature (Ivanov et al., 1978), under mild conditions these problems can be shown to be equivalent in a certain sense (Vasin, 1970; Bertsekas, 1999). Consequently the appropriate problem setting can be decided by the availability of computational tools to solve the respective optimization problems. In this thesis we consider problems of the first (unconstrained) and second (constrained) type.

1.2 Research Approach

Much research in structured sparsity centres on studying penalties Ω, primarily focusing on the structure they impose in the solution, their statistical properties, and methods to solve the resulting optimization problems. In particular, norms exhibit a number of properties which make them convenient choices for penalties, as we discuss in Chapter 2. Well known examples include ℓ_p-norms (Tibshirani, 1996; Zou and Hastie, 2005) and mixed $\ell_{p/q}$-norms (Yuan and Lin, 2006; Fornasier and Rauhut, 2008; Teschke and Ramlau, 2007; Kowalski, 2009) for sparse vector learning problems, as well as penalties for low rank matrix learning problems such as the trace norm (Srebro et al., 2005; Abernethy et al., 2009; Jaggi and Sulovsky, 2010; Mazumder et al., 2010). Other matrix penalties have been studied in multitask learning, where related learning problems are solved by taking advantage of commonalities between them (Evgeniou et al., 2005; Argyriou et al., 2007a, 2008; Jacob et al., 2009a).

Of particular relevance to this thesis are a range of norms that are defined using a variational formulation. These include a series of norms that are defined by an optimization over a set of atoms (Jacob et al., 2009b; Argyriou et al., 2012; Chandrasekaran et al., 2012b; Bach, 2013) or over a parameterized set (Micchelli and Pontil, 2005; Bach et al., 2012; Micchelli et al., 2013). While employing such penalties comes at a cost, such as intractability of computing the norms, the additional complexity of these penalties allows greater flexibility in encoding structural assumptions.

Following on from this line of research, in this thesis we study a number of norms that are defined via an optimization problem. Continuing a line of research by Micchelli and Pontil (2005); Bach et al. (2012), in Chapter 3 we first study a family of vector norms that are defined as an infimum of quadratics, parameterized over a convex set. Specifically, letting $\Theta \subset \mathbb{R}^d_{++}$ be a convex bounded subset of the positive orthant, we define the norm of $w \in \mathbb{R}^d$ by

$$\|w\|_\Theta = \left(\inf_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i} \right)^{\frac{1}{2}}.$$

In addition to the ℓ_p-norms, for $p \in [1, \infty]$, the Θ-norms capture the so-called Λ-norms for structured sparsity of Micchelli et al. (2013). As we show, a particular norm that is included in this family is the k-support norm of Argyriou et al. (2012). This was introduced as an alternative to the ℓ_1-norm of the lasso method and it was shown to perform well in sparse vector estimation problems. Similarly, the Moreau envelope of the k-support norm, which we term the box-norm, can also be written as a Θ-norm. Using this framework, we provide an efficient method to compute the proximity operator of the squared norm, which allows us to employ efficient optimization methods to solve regularization problems using the norms.

Related to the problem of learning a sparse vector is the problem of learning a low rank matrix, that is a matrix with a sparse spectrum. The trace norm, or nuclear norm, for matrices has been studied in the context of learning low rank matrices (Srebro et al., 2005; Abernethy et al., 2009; Jaggi and Sulovsky, 2010; Mazumder et al., 2010), and it is equivalent to ℓ_1-norm regularization of the spectrum of the learned matrix. This relationship is generalized by a classical theorem by von Neumann (Von Neumann, 1937), which provides that a vector norm that is a symmetric gauge function induces an orthogonally invariant matrix norm when applied to the spectrum of a matrix. It follows that under certain symmetry conditions on the parameter set Θ, the Θ-norm induces an orthogonally invariant matrix norm. Consequently, under appropriate conditions, a subset of the Θ-norms can be extended to matrices in this manner, and in Chapter 4 we use this fact to introduce the spectral k-support norm for low rank matrix learning. The norm generalizes in a natural manner the corresponding vector norm. Specifically, the unit ball of the k-support norm is the convex hull of vectors with cardinality k and unit ℓ_2-norm, and the unit ball of the spectral k-support norm is the convex hull of matrices with of rank k and unit Frobenius norm. Similarly, we obtain the spectral box-norm, and we show that the latter is closely related to the cluster norm of Jacob et al. (2009a), which has been used in multitask learning.

The k-support norm can naturally be extended in a number of ways. By generalizing the euclidean norm constraint in the unit ball definition to an ℓ_p-norm, for $p \in [1,\infty]$, in Chapter 5 we obtain the (k,p)-support norm, and the corresponding spectral (k,p)-support norm for matrices. For low rank matrix learning, the additional hyperparameter allows us to tune the decay of the spectrum of the underlying model, which can lead to better performance. We provide a Frank-Wolfe optimization method to solve constrained optimization problems with the norm when p is finite, and a method for the case $p = \infty$ using the projection onto the unit ball of the norm.

In recent years there has been increased interest in learning multilinear models. The entries of a tensor of dimension $p_1 \times \ldots \times p_m$ can be rearranged into m matrices, referred to as the matricizations of the tensor. A standard approach to regularization methods for tensors is to apply matrix regularizers to the matricizations of a tensor. In Chapter 6 we show that the natural extension of the k-support norm to tensors is the norm whose unit ball is the convex hull of the set of tensors whose matricizations have rank k and unit Frobenius norm. We further describe an optimization method and present numerical experiments using the norm.

In addition to the Θ-norm formulation, the k-support norm can be formulated as an infimal convolution (Rockafellar, 1970). In Chapter 7, we introduce a general optimal interpolation problem that defines a norm, and which captures many penalties used in machine learning, including the (k,p)-support norm. Specifically, for $w \in \mathbb{R}^d$, the norm $|||\cdot|||$ is defined as

$$|||w||| = \inf_{\substack{v \in \mathbb{R}^N \\ Bv = w}} \|F(v)\|, \tag{1.1}$$

where $\|\cdot\|$ is a monotone norm, F a vector-valued mapping the components of which are also norms, and B a $d \times N$ matrix. By choosing the constituent components appropriately we show that a number of regularizers belong to this family. In particular, when the set Θ has finite extreme points, the Θ-norm belongs to the class defined in (1.1). We then focus on the case where $\|\cdot\|$ is the ℓ_1-norm and describe an optimization algorithm which can be used to solve regularization problems involving the norms, and we present numerical experiments.

1.3 Contributions

In summary, the main contributions of this thesis are the following:

Θ-norms for sparse vector problems

In Chapter 3 we study the Θ-norms, which are defined as optimization problems over a convex set Θ.

- We show that this framework captures a number of norms, such as existing norms for structured sparsity, and the recently proposed k-support norm.
- We show that the vector k-support norm is a special case of the more general *box-norm*, which in turn can be seen as a perturbation of the former.
- The box-norm can also be written as a Θ-norm, and this framework is instrumental in deriving a fast algorithm to compute the norm and the proximity operator of the squared norm, which allows us to use optimal first order optimization algorithms for the box-norm.

Orthogonally invariant norms for low rank matrix problems

In Chapter 4 we discuss the application of Θ-norms to induce orthogonally invariant matrix norms for low rank matrix problems.

- We extend the k-support and box-norms to orthogonally invariant matrix norms. We note that the spectral box-norm is closely related to the cluster norm, which in turn can be interpreted as a perturbation of the spectral k-support norm.

- Our computation of the vector box-norm and its proximity operator extends naturally to the matrix case, which allows us to use proximal gradient methods for the cluster norm.
- We provide a method to apply the centered versions of the penalties, which are important in applications.
- We present numerical experiments on both synthetic and real matrix learning datasets.

The (k,p)-support norm

In Chapter 5 we introduce the (k,p)-support norm.

- We propose the (k,p)-support norm as an extension of the k-support norm and we characterize in particular the unit ball of the induced orthogonally invariant matrix norm.
- We show that the norm can be computed efficiently and we discuss the role of the parameter p.
- We outline a conditional gradient method to solve the associated regularization problem for both vector and matrix problems.
- In the special case $p=\infty$ we provide an $\mathcal{O}(d\log d)$ computation of the projection operator
- We present numerical experiments on matrix completion benchmarks that demonstrate that the proposed norm offers significant improvement over previous methods.

The tensor k-support norm

In Chapter 6 we extend the k-support norm to tensors.

- We define the tensor k-support norm and show that it arises naturally by considering the convex hull of the union of tensors whose matricizations are constrained in rank.
- We give a bound on the Rademacher average of the linear function class associated with the norm.
- We provide an optimization algorithm to solve the corresponding regularization problems and we provide numerical experiments.

Minimum interpolation norms

In Chapter 7 we study a general framework for a wide range of norms.

- We introduce a class of convex regularizers that encompasses many norms used in machine learning and statistics.

- We provide basic properties and examples of this construction, including the latent group lasso, the overlapping group lasso and the (k,p)-support norm.
- We present a general method to solve optimization problems involving regularizers of this form which is based on a recent stochastic block-coordinate Douglas-Rachford algorithm.
- We present numerical experiments using the latent group lasso and the non smooth hinge loss and we investigate the trade off between the number of block updates and total computation time.

1.4 List of Publications

The following publications were completed over the course of this thesis.

- Regularization with the Spectral k-Support Norm (with M. Pontil, D. Stamos, NIPS 2014)
- Fitting Spectral Decay with the k-Support Norm (with M. Pontil, D. Stamos, AISTATS 2016)
- New Perspectives on k-Support and Cluster Norms (with M. Pontil, D. Stamos, JMLR 2016)
- Learning with optimal interpolation norms: properties and algorithms (with C. L. Combettes, C. A. Micchelli, M. Pontil) arXiv:1603.09273, March 2016.

1.5 Outline

This thesis is organized as follows. In Chapter 2 we review literature related to structured sparsity regularization. In Chapter 3 we review the Θ-norms and we discuss the k-support norm. In Chapter 4 we discuss matrix regularizers and we introduce the spectral k-support norm. In Chapter 5 we introduce the (k,p)-support norm. In Chapter 6 we extend the k-support norm to tensors. In Chapter 7 we introduce the optimal interpolation norm framework. Finally, in Chapter 8 we discuss further work based on the topics covered in this thesis and we conclude. The appendices contain derivations of results that are not included in the body of the thesis.

Chapter 2

Background

A standard approach in machine learning and statistics is to constrain the space in which we seek the solution to an estimation problem in order to improve how well the model generalizes to unseen data. By choosing the constraint to encode certain characteristics that we assume to be present in the underlying model, the learned solution should more accurately reflect reality. An approach that has proven very successful in the last two decades consists of methods that promote sparsity. These assume that the data can be accurately described by a model that is primarily determined by a small number of parameters (Bach et al., 2012; Huang et al., 2011; Jenatton et al., 2011a; Maurer and Pontil, 2012; Hastie et al., 2015). Incorporating the assumption into the optimization problem should subsequently lead to solutions with the desired sparsity structure, providing improved predictive or descriptive performance.

Within this field, a particular line of research involves the study of penalties that scale with the complexity of the underlying object, that is, their value increases as sparsity of the object decreases. By incorporating these penalties into the learning problem, either directly via a constraint on the solution space, or indirectly via a trade off between model fit and sparsity, the solution is encouraged to be sparse, see e.g. Bach et al. (2011, 2012); Micchelli et al. (2013); Wainwright (2014). These methods have successfully been applied to a number of applications including computer vision (Huang et al., 2011; Jenatton et al., 2010), sparse coding (Jenatton et al., 2011b), collaborative filtering (Srebro et al., 2005; Abernethy et al., 2009), multitask learning (Evgeniou et al., 2005; Argyriou et al., 2007a, 2008; Jacob et al., 2009a; Obozinski et al., 2010), bioinformatics (Rapaport et al., 2008; Jacob et al., 2009b; Gramfort and Kowalski, 2009; Kim and Xing, 2010) and others.

In this review chapter, we first describe our general framework for structured sparsity problems. The learning task is framed as an optimization problem, and the concept of convexity is central to guarantee existence and uniqueness of solutions, as well as allowing access to the wealth of numerical methods to solve the corresponding optimization problem.

We then review core concepts from convex analysis and optimization that we will use throughout this thesis. Finally, we survey a number of structured sparsity penalties from the literature, with a particular focus on sparse vector and low rank matrix learning problems.

2.1 Problem Setting

Given data $\{(x_i,y_i)_{i=1}^n\} \subset \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ and $\mathcal{Y}$ are Euclidean spaces, we seek a model $f : \mathcal{X} \mapsto \mathcal{Y}$ in a finite dimensional Hilbert space $\mathcal{H}$ such that the model fits that data and generalizes well to unseen data. This is measured by $\mathcal{L} : \mathcal{H} \to \mathbb{R}$. In structured learning problems we introduce a functional $\Omega : \mathcal{H} \to \mathbb{R}_+$, such as a norm raised to some power, that measures the complexity (e.g. sparsity) of a model.

The manner in which the choice of Ω influences the structure of the underlying model depends on the specific framework that is chosen. For illustration we consider a simple linear regression task and we briefly review three problem settings from the inverse problems literature (Tikhonov and Arsenin, 1977). We are given data $X \in \mathcal{X}$ and observations $y \in \mathcal{Y}$ for which it holds

$$y = Xw^* + \varepsilon,$$

where in this case $X \in \mathbb{R}^{n\times d}$, $Y = \mathbb{R}^n$, $\mathcal{H} = \mathbb{R}^d$, w^* is the ground truth, and ε represents noise. We seek $w \in \mathcal{H}$ which minimizes $\mathcal{L}(w) = \|Xw - y\|$, where $\|\cdot\|$ is some norm, and we consider three optimization settings. With Tikhonov regularization, we solve the unconstrained problem

$$\inf_{w\in\mathcal{H}} \mathcal{L}(w) + \lambda\Omega(w), \tag{2.1}$$

where $\lambda > 0$ is the regularization parameter and regulates the trade-off between fitting the data and minimizing the complexity of the model. In contrast, Ivanov regularization involves a constrained problem of the form

$$\begin{aligned} &\inf_{w\in\mathcal{H}} \mathcal{L}(w) \\ &\text{s.t. } \Omega(w) \le \alpha, \end{aligned} \tag{2.2}$$

where $\alpha > 0$ determines the space within which we seek the model w. This setting can be interpreted as seeking the best fitting model with a given complexity defined by the level set $\{w \mid \Omega(w) \le \alpha\}$. Finally, Morozov regularization also involves a constrained problem, taking the form

$$\begin{aligned} &\inf_{w \in \mathcal{H}} \Omega(w) \\ &\text{s.t. } \mathcal{L}(w) \leq \beta, \end{aligned} \tag{2.3}$$

where $\beta > 0$ regulates the maximum error. This formulation can be interpreted as seeking the model with the smallest complexity, subject to a worst case error, as defined by the set $\{w \mid \mathcal{L}(w) \leq \beta\}$. The latter formulation in particular has been studied in the compressed sensing literature, with recent research being motivated by fundamental results by Donoho (2006); Candes et al. (2006) and others, see e.g. Qaisar et al. (2013); Hegde et al. (2014) for a review.

In a supervised learning setting, for each of the problems, the optimal value of λ, α and β can be selected by validation over the positive real numbers. In (2.1), the objective $\mathcal{L} + \lambda\Omega$ is composite and the penalty Ω is also referred to as the regularizer; in (2.2) and (2.3) we solve a constrained problem.

Under relatively weak assumptions, including convexity of $\mathcal{L}$ and Ω (see below), these problems are equivalent in the sense that the regularization paths, that is, the sets of solutions to each of (2.1), (2.2), and (2.3), are equivalent as we let λ, α and β vary over the positive reals, see e.g. Bach et al. (2011, §1.2), or Borwein and Lewis (2000, §4.3) for a discussion on the Tikhonov and Ivanov problems, or Vasin (1970) for the particular case of the Euclidean loss for all three problems.

Moreover, for a given regularizer Ω, the problems involving the regularizer or a power (greater than 1) are equivalent in the same sense. Specifically, for any $p, \tilde{p} > 1$, the Tikhonov problem with objective $\mathcal{L} + \lambda\Omega^p$ is equivalent to an Ivanov problem with constraint $\Omega^p \leq \alpha$, or by positivity, with constraint $\Omega^{\tilde{p}} \leq \tilde{\alpha}$, where $\tilde{\alpha} = \alpha^{\tilde{p}/p}$. Finally this is equivalent to a Tikhonov problem with objective $\mathcal{L} + \tilde{\lambda}\Omega^{\tilde{p}}$ (Bach et al., 2011, §1.2). A similar reasoning can be applied to the Ivanov and Morozov problems. This allows us to choose the power of the penalties according to the optimization problem that we are able to solve.

Much of the recent research in structured sparsity has involved studying penalty functionals that are proposed as the penalty Ω in problems (2.1), (2.2), and (2.3), in particular studying the structure of the sparsity pattern that they encourage, their relationship to other penalties, analysing their statistical properties, and deriving numerical methods to solve the corresponding optimization problems for one of the problem settings (Bach et al., 2012; Huang et al., 2011; Wainwright, 2014). In this thesis we continue this line of research, and we study a number of penalties for sparse learning problems, with a focus on the properties of the penalties, and optimization methods for the regularization problems. For the remainder of this chapter, we review fundamental concepts, principally from convex analysis and convex optimization, that we use throughout this thesis, and we review the literature for structured sparsity penalties.

2.2 Fundamentals

In this section we review our notation, we recall basic concepts from convex analysis and convex optimization, we discuss the existence of solutions to the optimization problems that we study throughout this thesis, and finally we recall an optimization method that can be used to solve a range of Tikhonov regularization problems.

2.2.1 Notation

We first outline common notation used throughout the thesis. Additional notation will be introduced in the following chapters as needed. We use $\mathbb{N}_d$ for the set of integers from 1 up to and including d. We let $\mathbb{R}^d$ be the d dimensional real vector space, whose elements are denoted by lower case letters. We let $\mathbb{R}^d_+$ and $\mathbb{R}^d_{++}$ be the subsets of vectors with nonnegative and strictly positive components, respectively. On $\mathbb{R}^d$, the canonical scalar product is given by $\langle w \mid u\rangle = \sum_{i=1}^d w_i u_i$, for $w, u \in \mathbb{R}^d$. For $p \in [1,\infty)$ the ℓ_p-norm of a vector $w \in \mathbb{R}^d$ is defined as $\|w\|_p = (\sum_{i=1}^d |w_i|^p)^{1/p}$ and $\|w\|_\infty = \max_{i=1}^d |w_i|$. The corresponding dual norm is $\|\cdot\|_{p,*} = \|\cdot\|_q$, where $1/p + 1/q = 1$. We denote by Δ^d the unit simplex in $\mathbb{R}^{d+1}$, $\Delta^d = \left\{\lambda \in \mathbb{R}^{d+1}_+ \mid \sum_{i=1}^{d+1} \lambda_i = 1\right\}$. For any vector $w \in \mathbb{R}^d$, its *support* is defined as supp$(w) = \left\{i \mid w_i \neq 0\right\} \subset \mathbb{N}_d$ and its *cardinality* is defined as the size of its support, card$(w) = \left|\left\{i \mid w_i \neq 0\right\}\right|$. We use 1 to denote either the scalar or a vector of all ones, whose dimension is determined by its context. Given $w \in \mathbb{R}^d$ and a subset g of $\mathbb{N}_d$, $w_{|g} \in \mathbb{R}^d$ is the vector whose components are identical to those of w on the support set defined by g, and zero elsewhere. In particular, the d-dimensional vector 1_g has ones on the support g, and zeros elsewhere. We let $\mathbb{R}^{d\times m}$ be the space of $d \times m$ real matrices and write $W = [w_1, \ldots, w_m]$ to denote the matrix whose columns are formed by the vectors $w_1, \ldots, w_m \in \mathbb{R}^d$. For a vector $\sigma \in \mathbb{R}^d$, we denote by diag(σ) the $d \times d$ diagonal matrix having elements σ_i on the diagonal. We say matrix $W \in \mathbb{R}^{d\times m}$ is *diagonal* if $W_{ij} = 0$ whenever $i \neq j$. We denote the *trace* of a matrix W by tr(W), its *transpose* by $W^\top$, and its *rank* by rank(W). The canonical inner product on $\mathbb{R}^{d\times m}$ is $\langle W \mid U\rangle = \mathrm{tr}(WU^\top)$, for $W, U \in \mathbb{R}^{d\times m}$. We use $\mathbf{S}^d$ to denote the set of real $d \times d$ symmetric matrices, and $\mathbf{S}^d_+$ to denote the subset of positive semidefinite matrices. We use $\succeq$ to denote the positive semidefinite ordering on $\mathbf{S}^d$. We use $\mathbf{O}^d$ to denote the set of real orthogonal $d \times d$ matrices. We let $\sigma(W) \in \mathbb{R}^r_+$ be the vector formed by the singular values of W, where $r = \min(d, m)$, and where we assume that the singular values are ordered nonincreasing, i.e. $\sigma_1(W) \geq \ldots \geq \sigma_r(W) \geq 0$. Similarly, for $W \in \mathbf{O}^d$, we let $\lambda(W)$ be the vector formed by the eigenvalues of W, counted by multiplicity, ordered nonincreasing. On

$\mathbb{R}^d$ we denote by $\|\cdot\|_2$ the *Euclidean norm*, and on $\mathbb{R}^{d\times m}$ we denote by $\|\cdot\|_{\mathrm{F}}$ the *Frobenius norm* and by $\|\cdot\|_{\mathrm{tr}}$ the *trace norm*, that is the sum of singular values, also known as the *nuclear norm*. The *convex hull* of a subset S of a vector space is denoted $\mathrm{conv}(S)$. Given a vector space $\mathcal{X}$, a subset $S \subset \mathcal{X}$, and function f defined on $\mathcal{X}$, a statement of the form "$f(S)$" should be taken as shorthand for "$f(s)$ *for all* s *in* S".

2.2.2 Basics

We now review some basic relevant definitions and results, see e.g. Bhatia (1997); Horn and Johnson (2005); Marshall and Olkin (1979). Throughout, let $\mathcal{X}$ be a Euclidean space.

Definition 1 (Norm). *A* norm $\|\cdot\|$ *on* $\mathcal{X}$ *is a function which is positive, homogeneous, non degenerate and satisfies the triangle equality, that is*

(i) $\|x\| \geq 0$, *for all* $x \in \mathcal{X}$,

(ii) $\|\alpha x\| = |\alpha|\|x\|$, *for all* $\alpha \in \mathbb{R}, x \in \mathcal{X}$,

(iii) $\|x\| = 0$ *if and only if* x *is the zero element in* $\mathcal{X}$,

(iv) $\|x+y\| \leq \|x\| + \|y\|$, *for all* $x, y \in \mathcal{X}$.

The corresponding dual norm is the function $\|\cdot\|_*$ *defined for* $y \in \mathcal{X}$ *by*

$$\|y\|_* = \sup\big\{\langle y \mid x\rangle \;\big|\; \|x\| \leq 1\big\},$$

where $\langle\cdot \mid \cdot\rangle$ *is the canonical scalar product on* $\mathcal{X}$.

Proposition 2 (Generalized Hölder inequality). *Let* $\|\cdot\|$ *be a norm on* $\mathcal{X}$. *Then, for all* $x, y \in \mathcal{X}$ *we have*

$$\langle x \mid y\rangle \leq \|x\|\|y\|_*.$$

The inequality follows immediately from the definition of the dual norm. The case of equality is of particular interest and we mention the following special cases.

Proposition 3 (Hölder Inequality). *Let* $x, y \in R^d$, *and let* $p, q \in (1,\infty)$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$. *Then*

$$\langle x \mid y\rangle \leq \|x\|_p\|y\|_q,$$

and equality holds when $x_i = \alpha sign(y_i)(|y_i|)^{\frac{q}{p}}$, *for any* $\alpha \in \mathbb{R}_+$. *Similarly*

$$\langle x \mid y\rangle \leq \|x\|_1\|y\|_\infty,$$

and equality holds when

$$x_i = \begin{cases} \alpha sign(y_i) & \text{if } i \in \mathcal{I} \\ 0 & \text{if } i \notin \mathcal{I} \end{cases},$$

where $\mathcal{I} = \{ i = 1 \ldots n \mid i = \operatorname{argmax}_{i=1}^{d} |y_i| \}$, *for any* $\alpha \in \mathbb{R}_+$.

Proof. The proof follows by direct computation, see Appendix A. □

The following notion is particularly useful in the context of regularizers for matrix learning problems, see e.g. Mirsky (1960).

Definition 4 (Symmetric gauge function). *A real-valued function* g*, defined on* $\mathbb{R}^d$*, is called a* symmetric gauge function *if it satisfies the following conditions.*

(i) $g(x) > 0 \, (x \neq 0)$;

(ii) $g(\alpha x) = |\alpha| g(x)$;

(iii) $g(x+y) \leq g(x) + g(y)$;

(iv) $g(Px) = g(x)$;

(v) $g(Jx) = g(x)$,

where P *is any permutation matrix, and* J *is any diagonal matrix with entries* ± 1.

The first three properties characterize a norm, hence it is clear that any norm on $\mathbb{R}^d$ which is invariant under permutations and sign changes is a symmetric gauge function, and vice versa. The following result is due to Von Neumann (1937), see also Lewis (1995).

Theorem 5 (OI norms and symmetric gauge functions). *A matrix norm is orthogonally invariant if and only if it is of the form* $\|X\| = g(\sigma(X))$*, where* $\sigma(X)$ *is the vector formed by the singular values of* X*, and* g *is a symmetric gauge function.*

The dual norm of an orthogonally invariant norm derives in the natural manner from the dual of the inducing symmetric gauge function.

Proposition 6. *Let* $\|\cdot\|$ *be an orthogonally invariant norm induced by the symmetric gauge function* g*. Then, for* $Y \in \mathbb{R}^{d \times m}$ *the dual norm is given by* $\|Y\|_* = g_*(\sigma(Y))$ *where* g_* *is the dual of* g*.*

Proof. See Appendix A. □

Theorem 7 (Von Neumann's trace inequality). *If $X, Y \in \mathbb{R}^{d \times m}$, then it holds*

$$\operatorname{tr}(XY^\top) \leq \langle \sigma(X) \mid \sigma(Y) \rangle$$

and equality holds if and only if X and Y admit a simultaneous singular value decomposition, that is, $X = U \operatorname{diag}(\sigma(X))V^\top$, $Y = U \operatorname{diag}(\sigma(Y))V^\top$, where $U \in \mathbf{O}^d$ and $V \in \mathbf{O}^m$ are orthogonal matrices.

Von Neumann's inequality provides an upper bound for the trace of AB. The following inequality provides a lower bound, and is given in Marshall and Olkin (1979, Sec. 9 H.1.h).

Lemma 8. *If $A, B \in \mathbf{S}^d_+$, then it holds*

$$\operatorname{tr}(AB) = \sum_{i=1}^{d} \lambda_i(AB) \geq \sum_{i=1}^{d} \lambda_i(A) \lambda_{d-i+1}(B).$$

A related result to Theorem 7 involving the spectral decomposition is known as Fan's inequality (Borwein and Lewis, 2000, Thm. 1.2.1).

Theorem 9 (Fan's inequality). *If $X, Y \in \mathbf{S}^d$, then it holds*

$$\operatorname{tr}(XY^\top) \leq \langle \lambda(X) \mid \lambda(Y) \rangle$$

and equality holds if and only if X and Y admit a simultaneous ordered spectral decomposition, that is, $X = U \operatorname{diag}(\sigma(X))U^\top$, $Y = U \operatorname{diag}(\sigma(Y))U^\top$, where $U \in \mathbf{O}^d$ is an orthogonal matrix.

Let $x, y \in \mathbb{R}^d$, and let $x^\downarrow$ and $y^\downarrow$ denote the vectors with the same components permuted into nonincreasing order, respectively. The following inequality is a direct consequence of Theorem 9, see also Borwein and Lewis (2000, Prop. 1.2.4).

Theorem 10 (Hardy-Littewood-Polya). *Any vectors $x, y \in \mathbb{R}^d$ satisfy the inequality*

$$\langle x \mid y \rangle \leq \left\langle x^\downarrow \mid y^\downarrow \right\rangle. \tag{2.4}$$

2.2.3 Convex analysis

Convexity is a crucial property in regularization penalties, as we discuss in this and the following section. The literature is vast, with core references including Bauschke and Combettes (2011); Bertsekas et al. (2003); Bertsekas (2015); Borwein and Lewis (2000); Boyd and Vandenberghe (2004); Hiriart-Urruty and Lemaréchal (2001); Rockafellar (1970). Throughout this section we consider learning problems in a real finite dimensional Hilbert space $\mathcal{X}$. Recall that the general problem is Tikhonov regularization, that is

$$\min_{x \in \mathcal{X}} \mathcal{L}(x) + \lambda \Omega(x), \tag{2.5}$$

where $\mathcal{L}$ is the loss function and Ω is a regularizer, typically a norm. Informally, convexity ensures that local minima are global minima, and we can employ descent based numerical methods to solve the learning problem. For the following fundamental definitions, we generally follow Hiriart-Urruty and Lemaréchal (2001); Borwein and Lewis (2000); Bertsekas (2009).

Definition 11. *A set C is* convex *if, for all x, y in C, for all λ in $[0,1]$, $\lambda x + (1-\lambda)y \in C$.*

Geometrically, any two points x and y in C can be linked by a segment entirely contained in C. In particular, note that $\mathbb{R}^d$ and $\mathbb{R}^{d\times m}$ are convex sets, as is the unit d-simplex Δ^d.

Definition 12. *A set $\Lambda \in \mathcal{X}$ is a* cone *if it satisfies $\alpha\Lambda = \Lambda$ for all $\alpha > 0$. If Λ is convex, we refer to it as a* convex cone.

Examples of convex cones include the set $\{x \in \mathbb{R}^d \mid \langle x \mid u \rangle \leq 0\}$, for a given $u \in \mathbb{R}^d$ and the (strictly) positive orthant $\mathbb{R}^d_{++}$.

Definition 13. *A* convex combination *of elements $\{x_1, \ldots, x_k\}$ of $\mathcal{X}$ is an element $\sum_{i=1}^k \alpha_i x_i$, where $\alpha_i \geq 0$ and $\sum_{i=1}^k \alpha_i = 1$.*

Definition 14. *Let C be a non empty convex set. The set of extreme points ext(C) is the set of points $x \in C$ such that if $x = \lambda y + (1-\lambda)z$, for $y, z \in \mathcal{X}$, and $\lambda \in (0,1)$, then $x = y = z$.*

Definition 15. *The* convex hull $\operatorname{conv} C$ *of a set C is equivalently defined as either of the following*

a) *the smallest convex set containing C, or*

b) *the intersection of all convex sets containing C.*

The second characterization holds as convexity is preserved by taking intersections.

Theorem 16 (Minkowski)**.** *Any compact convex set $C \subset \mathcal{X}$ is the convex hull of its extreme points.*

Note that when the set ext(C) has finitely many elements, C is a polyhedron. For the following 3 definitions let $C \subset \mathcal{X}$ be nonempty and convex.

Definition 17. *Let $x \in C$. The* normal cone *to C at x is $N_C(x) = \{u \in \mathcal{X} \mid \langle C - x \mid u \rangle \leq 0\}$.*

Definition 18. *The* polar set *of C is $C^{\ominus} = \{u \in \mathcal{X} \mid \langle C \mid u \rangle \leq 1\}$.*

Definition 19. *Let $x \in C$. The* tangent cone *to C at x is the polar of the normal cone, $T_C(x) = \{u \in \mathcal{X} \mid \langle N_C(x) \mid u\rangle \leq 0\}$.*

Having defined the geometric concept of convexity for sets, we can define the concept for functions.

Definition 20. *Let C be a convex set. A function $f : C \to \mathbb{R}$ is* convex *if for all x,y in C, and for all λ in $[0,1]$ we have $f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y)$. If the inequality is strict, then f is strictly convex.*

Definition 21. *A convex function f is* proper *if $f(x) < \infty$ for at least one x, and $f(x) > -\infty$ for all x.*

Definition 22. *The* domain *of $f : \mathcal{X} \to]-\infty,\infty]$ is the set dom $f = \{x \in \mathcal{X} \mid f(x) < \infty\}$.*

Definition 23. *The* epigraph *of a function $f : \mathcal{X} \to [-\infty,\infty]$ is the subset*

$$\mathit{epi} f = \{(x,\xi) \in \mathcal{X} \times \mathbb{R} \mid f(x) \leq \xi\}.$$

Definition 24. *Let $f : \mathcal{X} \to [-\infty,\infty]$ be convex. We say f is* closed *if its epigraph epi$f$ is closed.*

Related to closedness is the concept of *lower semicontinuity*.

Definition 25. *A function $f : \mathcal{X} \to [-\infty,\infty]$ is lower semicontinuous at an element $x \in \mathcal{X}$ if*

$$f(x) \leq \lim_{k\to\infty} \inf f(x_k)$$

for every sequence $\{x_k\} \subset \mathcal{X}$ with $x_k \to x$. We say that f is lower semicontinuous *if it is lower semicontinuous at each point x in its domain $\mathcal{X}$.*

Note that continuity implies lower semicontinuity. In the sequel, we let $\Gamma_0(\mathcal{X})$ be the set of proper lower semicontinuous convex functions from $\mathcal{X}$ to $]-\infty,\infty]$, and we use Γ_0 when there is no ambiguity about the domain. We have the following result (Bertsekas, 2009, Prop. 1.1.2).

Proposition 26. *For a function $f : \mathcal{X} \to [-\infty,\infty]$, the following are equivalent.*

a*) The level set $V_\gamma = \{x \mid f(x) \leq \gamma\}$ is closed for every scalar γ.*

b*) f is lower semicontinuous*

c*) epi(f) is closed.*

Definition 27. *A function $f : \mathcal{X} \to]-\infty,\infty]$ is* coercive *if*

$$\lim_{\|x\|\to\infty} f(x) = \infty.$$

The next concepts are fundamental auxiliary functions. Let $C \subset \mathcal{X}$.

Definition 28. *The* indicator function ι_C *of* C *is defined as*

$$\iota_C(x) = \begin{cases} 1, & \text{if } x \in C \\ 0, & \text{if } x \notin C. \end{cases}$$

Definition 29. *Let* C *be convex. The* characteristic function δ_C *of* C *is defined as* $\delta_C(x) = -\log I_C(x)$*, that is*

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

Definition 30. *The* support function σ_C *of* C *is defined as*

$$\sigma_C(x) = \sup \langle C \mid u \rangle.$$

Convexity is closely related to the triangle inequality for norms as the following two results show. The proofs follow directly from the definition of convexity and the properties of norms, see Appendix A.

Lemma 31. *A norm is convex and proper.*

We have the following converse.

Lemma 32. *Let* $f : \mathcal{X} \to \mathbb{R}$ *be positive, homogeneous and convex. Then* f *satisfies the triangle inequality. If* f *moreover is only 0 at the origin, then* f *is a norm.*

Convexity of a function provides a method to construct convex sets.

Lemma 33. *Let* C *be a convex set, and let* $f : C \to R$ *be convex. The level sets* $C_\alpha = \{x \in \mathcal{X} \mid f(x) \leq \alpha\}$ *are convex.*

The result is a direct consequence of the convexity of f. In particular, given a norm $\|\cdot\|$, for any $\alpha > 0$ the level set $\mathcal{B}_\alpha = \{x \in \mathcal{X} \mid \|x\| \leq \alpha\}$ corresponds to the unit ball of the norm scaled by α, and is a convex set. This fact ensures that the constraint sets in the Ivanov and Morozov problems are convex.

Convexity is trivially preserved through addition.

Lemma 34. *Let* $f, g : \mathcal{X} \to \mathbb{R}$ *be convex. Then* $f + g$ *is convex.*

Proof. The proof follows directly from the definition of convexity, see Appendix A. □

If either of the functions f or g is strictly convex then $(f+g)$ is strictly convex. Lemma 34 provides that when $\mathcal{L}$ and Ω are both convex, the objective in (2.5) is convex. Furthermore if $\mathcal{L}$ is a strictly convex loss function, such as the square loss, then the objective is strictly convex, which, under weak assumptions, guarantees uniqueness of solutions to (2.5). We return to this in Section 2.2.3.

Lemmas 31, 33 and 34 suggest that norms may be suitable candidates for penalties and we discuss this further in Section 2.3. In future chapters we shall make use of the following fundamental relationship between convex sets and norms in order to define new penalties.

Definition 35. *Let $C \subset \mathcal{X}$ be a closed convex set containing the origin. The* Minkowski functional *(or* gauge*) μ_C of C is defined, for every $x \in \mathcal{X}$, as*

$$\mu_C(x) = \inf\left\{\lambda \;\middle|\; \lambda > 0,\ \frac{1}{\lambda}x \in C\right\}.$$

The following result allows us to characterize the unit ball of a norm, see e.g. Rudin (1991, §1.35) for further details. Recall that a subset C of $\mathcal{X}$ is called *balanced* if $\alpha C \subseteq C$ whenever $|\alpha| \leq 1$. Furthermore, C is called *absorbing* if for any $x \in \mathcal{X}$, $x \in \lambda C$ for some $\lambda > 0$.

Lemma 36. *Let $C \subset \mathcal{X}$ be a bounded, convex, balanced, and absorbing set. The Minkowski functional μ_C of C is a norm on $\mathcal{X}$.*

Proof. We show that μ_C satisfies the properties of a norm, see Appendix A. □

Note that for such set C, the unit ball of the induced norm μ_C is C. Furthermore, if $\|\cdot\|$ is a norm then its unit ball satisfies the hypotheses of Lemma 36.

The Fenchel conjugate is a fundamental transform in convex analysis and plays a role similar to the gradient of a smooth function, see e.g. Hiriart-Urruty and Lemaréchal (2001, §E) for a discussion.

Definition 37 (Fenchel conjugate)**.** *The* conjugate (or Fenchel conjugate) *of a function $f : \mathcal{X} \to [-\infty, \infty]$ is the function $f^* : \mathcal{X} \to [-\infty, \infty]$ defined by*

$$f^*(u) = \sup_{x \in \mathcal{X}} \{\langle u \mid x\rangle - f(x)\}.$$

The function f^* is convex, and we define the biconjugate f^{**} as the conjugate of the conjugate f^*.

Example 38. *Examples of conjugates.*

a*) Let $C \subset \mathcal{X}$ and let $f = \delta_C$. Then $f^* = \sigma_C$, that is, the conjugate is the support function.*

b*) Let f be a norm $\|\cdot\|$. Then $f^* = \delta_{\mathcal{B}_*}$, where $\mathcal{B}_* = \{u \in \mathcal{X} \mid \|u\|_* \leq 1\}$, that is, the conjugate is the characteristic function of the unit ball of the dual norm.*

c) *Let* $f = \frac{1}{2}\|\cdot\|^2$. *Then* $f^* = \frac{1}{2}\|\cdot\|_*^2$.

Proof. See Appendix A. □

In general the penalty function Ω need not be smooth. Indeed, in general norms are not differentiable at the origin, hence in both the Tikhonov and Morozov settings, we require a more general concept than the gradient for smooth functions. The *subgradient* performs this role, and it is used to qualify the optimality conditions in non smooth optimization.

Definition 39 (Subgradient). *Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function. We say $u \in \mathcal{X}$ is a* subgradient *of f at a point $x \in \mathcal{X}$ if $f(x) + \langle u \mid y - x\rangle \leq f(y)$, for all $y \in \mathcal{X}$.*

If f is differentiable at x, then u is unique. Informally, a subgradient at x is a hyperplane which passes through x and which lower bounds $f(x)$.

Definition 40 (Subdifferential). *The set of all subgradients of a convex function f at $x \in \mathcal{X}$ is called the* subdifferential *of f at x, and is denoted by $\partial f(x)$, that is*

$$\partial f(x) = \big\{u \in \mathcal{X} \;\big|\; f(x) + \langle u \mid y - x\rangle \leq f(y), \textit{ for all } y \in \mathcal{X}\big\}.$$

If f is differentiable at x, the subdifferential reduces to the singleton containing the gradient of f at x, $\partial f(x) = \{\nabla f(x)\}$. We distinguish between strong subdifferential calculus in optimization theory, where the set of all subgradients at a point is required, and weak differential calculus in optimization applications, where identification of a single subgradient at a point is required.

The following result (Hiriart-Urruty and Lemaréchal, 1993, Ch. VI., Th. 2.2.1) characterizes the optimality conditions of a non smooth optimization problem in terms of the subdifferential of the objective function.

Theorem 41 (Optimality conditions). *For $f \in \Gamma_0$, the following two properties are equivalent.*

- *f is minimized at x^* over $\mathcal{X}$, that is $f(y) \geq f(x^*)$ for all $y \in \mathcal{X}$,*
- $0 \in \partial f(x^*)$.

When the function f is smooth, Theorem 41 reduces to the standard condition that the gradient vanishes, that is $\nabla f(x^*) = 0$.

A special case that we encounter throughout this thesis relates to linear objectives (Bertsekas, 2009, Prop. 2.4.2).

Proposition 42 (Supremum of linear functional over compact set). *Let P be a polyhedral set that has at least one extreme point. A linear function that is bounded below over P attains a minimum at some extreme point of P.*

We now introduce the proximity operator $\operatorname{prox} f$ of a convex function, which plays a key role in the optimization algorithms that we employ throughout this thesis (Moreau, 1965; Combettes, 2004; Parikh and Boyd, 2013; Villa, 2014). Let $\|\cdot\|$ be the norm induced by the standard inner product on $\mathcal{X}$.

Definition 43. *Let $f \in \Gamma_0$. The* proximity operator *of f at x with parameter ρ is defined as*

$$\operatorname{prox}_f(x) = \operatorname*{arginf}_{u \in \mathcal{X}} f(u) + \frac{1}{2\rho}\|x-u\|^2. \tag{2.6}$$

The objective in (2.6) is strictly convex, hence $\operatorname{prox}_f(x)$ is uniquely defined for each $x \in \mathcal{X}$. Note that $\operatorname{prox}_f(x)$ will only evaluate to x if it is an infimizer of f, otherwise $u = \operatorname{prox}(x)$ will be a point which trades off minimizing f and the distance from x. A necessary condition for u to be an infimizer of lower semicontinuous, proper, h is that $0 \in \partial h(x)$. We therefore obtain the following characterization of the proximity operator.

$$u = \operatorname{prox}_f(x) \Leftrightarrow 0 \in \partial f(x) + (x-u) \Leftrightarrow u - x \in \partial f(x) \Leftrightarrow u = (I + \partial f)^{-1}(x). \tag{2.7}$$

The Moreau envelope of f, which is given by the pointwise evaluation of the proximity operator, is convex and smooth (Moreau, 1965; Bauschke and Combettes, 2011). It acts as a parameterized smooth approximation to f from below, and it is an important tool in variational analysis (Rockafellar and Wets, 2009, Ch. 1 §G).

Definition 44. *Let $f \in \Gamma_0$. The* Moreau envelope *of f with parameter ρ is defined as*

$$e_\rho f(x) = \inf_{u \in \mathcal{X}} f(u) + \frac{1}{2\rho}\|x-u\|^2. \tag{2.8}$$

The following fundamental identity relates the proximity operator and Fenchel duality, see e.g. Parikh and Boyd (2013).

Theorem 45 (Moreau decomposition)**.** *Let $f \in \Gamma_0$. Then*

$$x = \operatorname{prox}_f(x) + \operatorname{prox}_{f^*}(x).$$

A direct consequence is the duality relationship

$$u = \operatorname{prox}_f(x) \Leftrightarrow x - u \in \partial f(u) \Leftrightarrow u \in \partial f^*(x-u) \Leftrightarrow x - (x-u) \in \partial f^*(x-u) \Leftrightarrow x - u = \operatorname{prox}_{f^*}(x).$$

Theorem 45 allows us to compute the proximity operator of a function whenever the proximity operator of its conjugate is known. Note that if $f = \delta_C$, where C is a non empty closed set, then the proximity operator reduces to the Euclidean projection Π_C onto C, as

$$\operatorname{prox}(x) = \operatorname*{argmin}_{x \in \mathcal{X}} \delta_C(u) + \frac{1}{2}\|u-x\|^2 = \operatorname*{argmin}_{x \in C} \|u-x\| = \Pi_C(x). \tag{2.9}$$

The case where f is a norm $\|\cdot\|$ is of particular interest. Recall first that the conjugate f^* is given by the characteristic function of the unit ball $\mathcal{B}_*$ of the dual norm, that is $f^*(u) = \delta_{\{\|\cdot\|_* \leq 1\}}(u)$. Moreau's decomposition then provides that the proximity operator of a norm can be computed via the projection on the unit dual norm ball, that is

$$\operatorname{prox}(x) = x - \Pi_{\mathcal{B}_*}(x).$$

2.2.4 Optimization

We can now characterize the existence of solutions to the optimization problems (2.1), (2.2), and (2.3). Our treatment follows Bertsekas (2009, §3.1). Consider the general problem

$$\inf_{x \in \mathcal{X}} f(x), \tag{2.10}$$

where $X \subset \mathcal{H}$. For Tikhonov regularization $\mathcal{X}$ corresponds to the entire space $\mathcal{H}$, with $f = \mathcal{L} + \lambda\Omega$. The Ivanov and Morozov problems correspond to $f = \mathcal{L}$ with $\mathcal{X} = \{\Omega(x) \leq \alpha\}$, respectively $f = \Omega$ with $\mathcal{X} = \{\mathcal{L}(x) \leq \beta\}$.

An element $x \in \mathcal{X} \cap \operatorname{dom}(f)$ is a *feasible* solution of the problem (2.10), and the problem is feasible if at least one feasible solution exists. Consequently we only consider $x \in X \cap \operatorname{dom}(f)$ as candidate solutions to (2.10). We note that with this restriction, the Tikhonov problem for a general objective is implicitly constrained. However, for machine learning problems that we consider in practice, for loss functions such as the square loss, and penalties defined by norms, the domain in each case is the entire space, hence the Tikhonov problem is not constrained.

The problem we seek to solve is to find $x^* \in \mathcal{X} \cap \operatorname{dom}(f)$ such that $f(x^*) = \inf_{x \in \mathcal{X}} f(x)$, and we call x^* a (global) minimizer. Equivalently, we have $x^* \in \operatorname{argmin}_{x \in \mathcal{X}} f(x)$. When the solution is unique, we may write $x^* = \operatorname{argmin}_{x \in \mathcal{X}} f(x)$ with a slight abuse of notation. If an element $\tilde{x}$ satisfies $f(\tilde{x}) = \inf\{f(x) \mid x \in \mathcal{X}, \|x - \tilde{x}\| \leq \epsilon\}$, for some $\epsilon > 0$, then we say $\tilde{x}$ is a local minimizer of (2.10). A fundamental consequence of convexity is that local minima are global minima. We summarize this as follows (Bertsekas, 2009, Prop. 3.1.1).

Proposition 46. *If X is a convex subset of $\mathcal{H}$ and $f : \mathcal{X} \to]-\infty, \infty]$ is a convex function, then a local minimum of f over $\mathcal{X}$ is also a global minimum. If in addition f is strictly convex, then there exists at most one global minimum of f over $\mathcal{X}$.*

We can now state the following corollary of the classical Weierstrass theorem (Bertsekas, 2009, Prop. A.2.7) which guarantees existence of optimal solutions (Bertsekas, 2009, Prop. 3.2.1).

Proposition 47 (Weierstrass). *Consider a closed proper function $f : \mathcal{X} \to]-\infty, \infty]$, and assume that any one of the following three conditions holds:*

a) *dom(f) is bounded.*

b) *There exists a scalar γ such that the level set $\{x \mid f(x) \leq \gamma\}$ is nonempty and bounded.*

c) *f is coercive.*

Then the set of minima of f over $\mathcal{X}$ is nonempty and compact.

For our purposes we obtain the following consequence of Proposition 47, which applies to the regularization problems of the form discussion above.

Proposition 48. *Let $\mathcal{H}$ be a finite dimensional Euclidean space, let $\mathcal{L}$ be convex, proper, and coercive on $\mathcal{H}$, and let Ω be a norm on $\mathcal{H}$. Let $\lambda, \alpha, \beta \in \mathbb{R}_+$. In each of the problems* (2.1), (2.2), *and* (2.3) *the set of solutions is nonempty and compact.*

Proof. The proof follows from Proposition 47, see Appendix A. □

2.2.5 Proximal gradient descent

While Proposition 48 establishes existence of global optima for the type of learning problems that we consider in this thesis, unfortunately in general computing the subdifferential of the objective function is not straightforward, hence we resort to numerical methods rather than applying Theorem 41 directly.

In this section we review a class of numerical algorithms which can be used to solve a wide range of regularization problems in machine learning with fast convergence rates. Specifically we consider problems of the form

$$\min_{x \in \mathcal{H}} f(x) + \lambda g(x), \tag{2.11}$$

where f is smooth with Lipschitz continuous gradient, g is convex and $\lambda > 0$ is a regularization parameter. As the form of (2.11) suggests, when f is a smooth loss function such as the square loss, and g is some regularizer Ω, the algorithm conveniently captures Tikhonov regularization problems in a natural manner. Moreover, choosing $\lambda > 0$ and letting $g(x) = \delta_{\{\Omega(\cdot) \leq \alpha\}}(x)$, for some norm Ω, and for $\alpha > 0$, we note that (2.11) also captures constrained problems of Ivanov type (2.2). A similar transformation is possible for problems of Morozov type (2.3).

The general algorithm requires, at each step, the computation of the gradient of the smooth term, ∇f, and the computation of the proximity operator of the penalty term, prox_g.

We now outline the iterative shrinkage thresholding algorithm (ISTA), following e.g. Beck and Teboulle (2009). For ease of notation we absorb the regularization parameter into the penalty, and we consider the problem

$$\min_{x\in\mathcal{H}} F(x), \tag{2.12}$$

where $F(x) = f(x) + g(x)$, and $\mathcal{H}$ is a Euclidean space with norm $\|\cdot\|$. We assume f is smooth, convex, with Lipschitz continuous gradient $L(f)$, so it holds

$$\|\nabla f(x) - \nabla f(y)\| \leq L(f)\|x - y\|, \quad \text{for every } x, y \in \mathcal{H},$$

and we assume g is continuous and convex. We purposefully do not require smoothness of g, which will allow the algorithm to be used with non smooth penalties as discussed in the following section. Finally, we assume that a solution x^* to (2.12) exists, and define $F^* = f(x^*)$. For any $L > 0$, in the first instance, we compute the quadratic approximation of $F(x) = f(x) + g(x)$ at point y:

$$Q_L(x, y) = f(y) + \langle x - y \mid \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2 + g(x).$$

This quantity admits a unique minimizer

$$\begin{aligned} p_L(y) &= \operatorname*{argmin}_x Q_L(x, y) \\ &= \operatorname*{argmin}_x g(x) + \langle x \mid \nabla f(y)\rangle + \frac{L}{2}\|x - y\|^2 \\ &= \operatorname*{argmin}_x g(x) + \frac{L}{2}\|x - (y - \frac{1}{L}\nabla f(y))\|^2 \\ &= \operatorname{prox}_{\frac{1}{L}g}(y - \frac{1}{L}\nabla f(y)) \end{aligned}$$

where we have dropped constant terms in y and completed the square. We obtain the iterative scheme $x^t = \operatorname{prox}_{\alpha g}(x^{t-1})$, where in this case $\alpha = \frac{1}{L}$. The method is outlined as the proximal gradient descent algorithm in Algorithm 1, also known as the iterative shrinkage-thresholding algorithm (ISTA). The term *shrinkage* derives from the fact that the prox is non expansive, that is, for $f \in \Gamma_0$, and $x, y \in \mathcal{H}$ it holds

$$\|\operatorname{prox}_f(y) - \operatorname{prox}_f(x)\| \leq \|y - x\|,$$

see e.g. Bauschke and Combettes (2011). The update involves computing the linear approximation to the smooth objective, and computing a regularization of this quantity via the proximity operator. At each step it requires a gradient (forward) computation and a proximity

Algorithm 1 ISTA.

```
Choose $\alpha > 0$. Choose $x^0 \in \mathcal{H}$.
for $t = 0, \ldots, T$ do
    Update $x^t := \mathrm{prox}_{\alpha g}(\nabla f(x^{(t)}))$
end for
```

operator (backward) computation, and the algorithm is also referred to as forward backward splitting by Combettes and Wajs (2005) or SpaRSA by Wright et al. (2009). The proximity operator can be interpreted as a general projection operator (Parikh and Boyd, 2013), hence the proximal gradient descent algorithms can be seen as a special case of projected gradient methods, see e.g. Bertsekas (2015). If $\alpha \leq \frac{1}{L}$ the objective in (2.12) decreases monotonically, that is, the algorithm converges monotonically. Furthermore, it has convergence speed $\mathcal{O}(1/t)$, that is, $F(x^t) - F(x^*) \approx \mathcal{O}(1/t)$, also known as *sublinear convergence* (Beck and Teboulle, 2009). An accelerated version of the algorithm with convergence speed $\mathcal{O}(1/t^2)$, but without monotone convergence, known as FISTA, involves computing an intermediate point y^t, in which case the update equation evaluates x^{t+1} as a convex combination of y^t and the proximity point, see e.g. Beck and Teboulle (2009) for full details. Alternative accelerated versions have also been proposed by Bioucas-Dias and Figueiredo (2007) and Nesterov (2007). Note that the proximal gradient descent algorithm reduces to a gradient descent method when $g = 0$ since the proximity operator of the zero operator is simply the identity operator.

In the general case, the proximal gradient algorithm requires the computation of the gradient of f and the proximity operator of g at each iteration, however the latter is itself an optimization problem. It follows that the algorithms are useful in those cases where the proximity operator is easy to compute, such as when it has a closed form expression, or there exists a computationally efficient procedure to compute the proximity operator. This is the case for a number of well known regularizers that we review in the following section, in particular the ℓ_1-norm of the Lasso method. In this thesis we make use of two further algorithmic frameworks, and we defer the discussion to later chapters. In particular, in Chapter 5 we discuss a conditional gradient method, which we apply to a constrained (Ivanov) regularization problem. Moreover, for Tikhonov regularization problems where neither the loss nor the penalty is smooth, we require a different approach to solving the problems, and in Chapter 7 we introduce a random sweeping splitting algorithm that addresses this case.

2.3 Norms as Penalties

In Section 2.2.3 we established that a suitable candidate for Ω should be continuous, convex and coercive. It follows that norms are natural candidates for penalties, and indeed much of the literature, and the entirety of this thesis, focuses penalties of this type.

Given a positive, convex penalty Ω, the level sets $\{w \mid \Omega(w) \leq \alpha\}$, for $\alpha > 0$, can provide insight into structural or statistical properties that Ω promotes. Specifically, in the case of a norm, the level sets correspond to the unit ball of the norm, scaled by α, which suggests a link between the geometry of the norm and the properties of the solution to a learning problem.

Geometric approaches to machine learning problems form an entire research area in itself, and a review of the literature is outside of the scope of this thesis, however we briefly mention a few results notions which are relevant to structured sparsity, in particular relating to aspects of optimization and statistical recovery properties.

In the constrained problem (2.2) in $\mathbb{R}^d$, let $\mathcal{B}$ be the ball of a norm penalty of radius α. At optimality the gradient of $\mathcal{L}$ evaluated at a solution $\hat{x}$ is contained within the normal cone of the ball of radius α at $\hat{x}$, that is the set $\{z \in \mathbb{R}^d \mid \langle x \mid z \rangle \leq 0\}$, see e.g. Bach et al. (2012, §1.3) and Borwein and Lewis (2000, §3.2). It follows that when α is small enough that the constraint is active, the level set of $\mathcal{L}$ for the value $\mathcal{L}(\hat{x})$ is tangent to the unit ball. If we inspect the ℓ_2-norm and ℓ_1-norm balls (see e.g. Figure 2.1 below, or 3.2 in Chapter 3), it is clear that the tangent sets at points located on their respective surfaces behave differently. More precisely, the smooth isotropic ℓ_2-norm unit ball treats all directions equally, whereas the anisotropic ℓ_1-norm ball has singularities located on the axes, which are specifically those points which have some zero coordinates, that is they correspond to sparse vectors.

Drusvyatskiy et al. (2015) consider the extreme points of the level set of a composite regularizer, that is the unit ball of a penalty which is defined as the sum of two distinct norms, for matrix learning. Penalties of this type have been studied in the literature in order to encode more than one constraint, such as matrices which are both sparse and low rank (Candes et al., 2011; Chandrasekaran et al., 2012a; Oymak et al., 2015). Drusvyatskiy et al. (2015) provide an algebra relating the respective unit balls, and use this to tune the constituent penalties.

In a related line of work, Amelunxen et al. (2014) link the geometry of a penalty Ω to the statistical recovery properties in sparse estimation problems in the Morozov regularization setting, extending work by Chandrasekaran et al. (2012b). The authors define the notional of *statistical dimension* of a convex cone, which, in a certain sense, measures its width. Using this quantity they relate the statistical recovery properties of a penalty Ω for a sparse convex problem to the shape of the descent cone of Ω at the optimal solution. As in the previous analysis, the sharp extreme points of the ℓ_1-norm ball contrast with the uniformly smooth surface of the ℓ_2-norm, and Amelunxen et al. (2014) show that this leads to better statistical recovery guarantees for the former.

In the following sections we survey a number of regularizers that have been proposed in the literature for structured sparsity. For a review of this material, see Huang et al. (2011);

Bach et al. (2012); Jenatton et al. (2011a); Hastie et al. (2015).

2.4 Vector Penalties

In the context of vector estimation, sparsity refers to the presence of zero components. The typical problem setting is linear regression with the square loss, however other settings have been explored, such as classification with the hinge loss, see e.g. Hastie et al. (2011), or ℓ_1-regularized logistic regression, see e.g. Krishnapuram et al. (2005). Sparse vectors are particularly useful in feature selection, as the features that have the most predictive power should correspond to non zero components of the regression vector, that is, the components in the support set of the vector.

In the ordinary least squares problem we have a sample of n points $\{(x_i, y_i)\}_{i=1}^n$, where $(x_i, y_i) \in \mathbb{R}^{d-1} \times \mathbb{R}$, and we solve

$$\min_{w \in \mathbb{R}^{d-1}, b \in \mathbb{R}} \|Xw + b1 - y\|_2^2 = \min_{w \in \mathbb{R}^{d-1}, b \in \mathbb{R}} \sum_{i=1}^n (x_i^\top w + b - y_i)^2. \tag{2.13}$$

If the data are not assumed to be centered, we can apply a standard technique to implicitly include the bias term in the data. Specifically, letting the data $x_i^\top$ form the rows of $\tilde{X} \in \mathbb{R}^{n \times (d-1)}$, suppose $(\tilde{w}, \tilde{b}) \in \mathbb{R}^{d-1} \times \mathbb{R}$ solve (2.13). Then letting $X = [\tilde{X}, 1] \in \mathbb{R}^{n \times d}$, the solution to the problem

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 = \min_{w \in \mathbb{R}^d} \sum_{i=1}^n (x_i^\top w - y_i)^2, \tag{2.14}$$

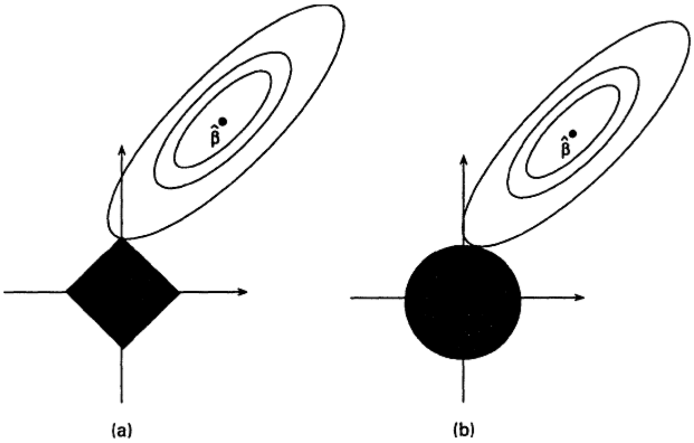
will be of the form $w^\top = [\tilde{w}, \tilde{b}]$, that is, we recover the regression vector and bias for problem (2.13). Unless otherwise stated, we will use the form of (2.14) going forward.

A number of norms have been proposed as regularizers to encourage structure in the regression vector. The standard ridge regression method uses the squared ℓ_2-norm, that is

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda \|w\|_2^2. \tag{2.15}$$

This method, introduced by Hoerl and Kennard (1970), guarantees that the problem admits a unique solution for a fixed λ, which is obtained by solving a linear system of equations. The effect of the ℓ_2-norm is to moderate fluctuations around the mean components of w, however these do not produce zero entries in general, so the resulting model is not sparse.

Figure 2.1: Geometry of lasso (left) and ridge regression (right) (Tibshirani, 1996).



2.4.1 Lasso

In order to explicitly control for the number of zeros in the regression vector and obtain sparse solutions, the problem we would like to solve is

$$\min_{w\in\mathbb{R}^d} \|Xw-y\|_2^2+\lambda\,\mathrm{card}(w), \tag{2.16}$$

where the cardinality operator card counts the number of non zero components of a vector. Unfortunately card is not convex since, for example, $\mathrm{card}(\frac{1}{2}(0+e_i)) > \frac{1}{2}\mathrm{card}(0)+\frac{1}{2}\mathrm{card}(e_i)$, for any standard basis vector e_i. Moreover, the regularization problem with the square loss and the card operator is NP-hard, see e.g. Natarajan (1995). Consequently, in order to employ tools from convex analysis, a standard approach is to use convex surrogates, the principal of which is the ℓ_1-norm, and various norms which are derived from it, as we now discuss.

With the square loss, regularization with the ℓ_1-norm is known as the lasso, for least absolute shrinkage and selection operator

$$\min_{w\in\mathbb{R}^d} \|Xw-y\|_2^2+\lambda\|w\|_1. \tag{2.17}$$

Tibshirani (1996) proposed using the ℓ_1-norm in the Ivanov setting, and provided the lasso name. Previously, Breiman (1995) had proposed a *nonnegative garotte*, involving a constrained ℓ_1-norm with a nonnegativity constraint on the regression vector components and Frank and Friedman (1993) suggested bounding the ℓ_p-norm, for $p>0$. In Tikhonov form the method can efficiently be solved using proximal gradient methods, which we outlined in Section 2.2.5. The effect of the ℓ_1-norm is to produce a regression vector w whose sparsity is determined by the magnitude of the regularization parameter.

Some intuition for using the ℓ_1-norm is given by the geometry of the unit ball of the norm.

Framing each of the ridge regression and lasso methods in their constrained formulation, the problem involves finding a solution at the intersection of an ellipsoid (the square loss) and the respective scaled unit balls of the norm. In contrast to the isotropic shape of the Euclidean norm ball, the sharper corners of the ℓ_1-norm ball encourage the intersection to occur at a point along the axes, that is, at a point where one or more coordinates are zero, especially in higher dimensions. Figure 2.1, taken from Tibshirani (1996), depicts this.

A more principled justification for the ℓ_1-norm comes from the fact that it is the tightest convex approximation to the card operator on the ℓ_∞-norm unit ball, that is, also known as the convex envelope, see e.g. Boyd and Vandenberghe (2004); Fazel (2002); Jojic et al. (2011). Figure 2.2 illustrates this relationship.

Following the success of the lasso method, a number of extensions have been proposed which build upon the ℓ_1-norm, and most of the penalties that we review in the following sections derive from it in a natural manner.

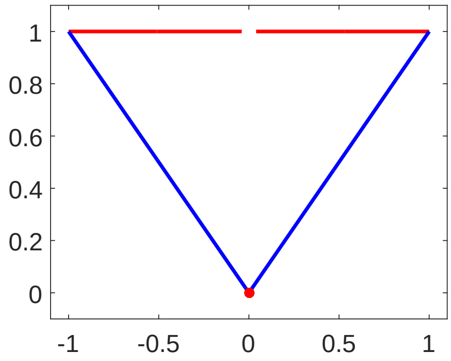
2.4.2 Elastic net

A common theme in regularizers for structured sparsity is combining simple penalties to achieve a combination of their effects, for example, controlling for sparsity and extreme values when combining the ℓ_1 and ℓ_2-norms. A natural extension to the lasso which takes this approach is the elastic net (Zou and Hastie, 2005). The model is

$$\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2, \tag{2.18}$$

where λ_1 and λ_2 are regularization parameters. This results in sparse regression vectors, whose non zero components are of similar size, that is it promotes a grouping effect where highly correlated features tend to be selected or excluded as a group. In particular in the high dimensional case where $d \gg n$, Zou and Hastie (2005) showed that the elastic net may outperform the lasso. While intuitively, a composite penalty imposes a combination of the effects of the component regularizers, each additional regularizer introduces an additional parameter, which increases the computational load of validation exponentially. Solving the optimization problem involving a composite regularizer is also typically less straightforward, for example when using methods involving the proximity operator, which we reviewed in Section 2.2.3. For the elastic net specifically, Zou and Hastie (2005) show that the elastic net problem can be expressed as a lasso problem on an augmented dataset, hence the same numerical methods can be applied, however in general this is not the case for other composite regularizers.

Figure 2.2: Geometric interpretation of ℓ_1-norm as convex envelope of card on ℓ_∞-norm unit ball.



2.4.3 Group lasso

The effect of the ℓ_1-norm penalty in the lasso is to select a subset of the support $\mathrm{supp}(w) \subset \mathbb{N}_d$ of the regression vector to be activated, that is the non zero coordinates. While the support of the regression vector found via the lasso tends to be sparse, the location of the non zero components is not known in advance. The group lasso is a method that was introduced to incorporate a priori assumptions on the relationship between the features in the data, or equivalently between the components of the regression vector (Yuan and Lin, 2006). Let $G = \{g_1, \dots, g_n\}$ be a subset of the power set of $\mathbb{N}_d$, such that the groups form a partition of $\mathbb{N}_d$, that is, let the groups be disjoint and let their union be $\mathbb{N}_d$. The group lasso norm is defined, for $w \in \mathbb{R}^d$, as

$$\|w\|_{\mathrm{GL}} = \sum_{i=1}^{n} \sqrt{p_i} \|w_{g_i}\|_2, \tag{2.19}$$

where p_i is a strictly positive weight, e.g. the size of g_i. The penalty aggregates the ℓ_2-norms of the vectors formed by the vector w restricted to each of the support sets g_i. The effect of the group lasso is to activate all of the components of the regression vector in each group as a whole, or not at all. Similar to the elastic net, within each group, extreme values, that is, values with large deviation from the group mean, are discouraged as the ℓ_2-norm acts separately on each subset w_{g_i}. The constrained method with a general loss function was introduced as blockwise sparse regression by Kim et al. (2006).

2.4.4 Sparse group lasso

The ℓ_2-norm applied to each group in (2.19) encourages the components of w_{g_i} to be of similar orders of magnitude. The sparse group lasso combines the group lasso with an additional ℓ_1-norm penalty in order to promote sparsity within the groups as well as with respect to the groups (Simon et al., 2013). The penalty is given, for $w \in \mathbb{R}^d$, by

$$\|w\|_{\text{SGL}} = \lambda_1 \sum_{i=1}^{n} \sqrt{p_i} \|w_{g_i}\|_2 + \lambda_2 \|w\|_1, \tag{2.20}$$

where λ_1, λ_2 are regularization parameters which regulate the relative importance of group-wide and within-group sparsity respectively.

2.4.5 Fused lasso

A modification of the lasso was proposed by Tibshirani and Saunders (2005). The method was motivated in particular by problems where the number of features is significantly greater than the number of data points. The fused lasso involves an ℓ_1-norm constraint, as well as a constraint on the difference in absolute value of successive components of the regression vector.

$$\begin{aligned} &\min_{w \in \mathbb{R}^d} \|Xw - y\|_2^2, \qquad (2.21)\\ &\text{s.t. } \|w\|_1 \leq \alpha_1, \sum_{i=2}^{d} |w_i - w_{i-1}| \leq \alpha_2, \end{aligned}$$

where α_1 and α_2 are positive parameters.

2.4.6 Composite absolute penalty

The group lasso penalty can be considered an ℓ_1/ℓ_2-norm, as it computes the sum of ℓ_2-norms. More generally it is an example of a mixed $\ell_p/\ell_{p'}$-norm, where $p, p' \in [1, \infty]$. The CAP penalty is an extension of the group lasso penalty which uses mixed norms (Zhao et al., 2009). Specifically, for $G = \{g_1, \ldots, g_n\}$, $p_0, \ldots, p_n \in [1, \infty]$, the penalty is given, for $w \in \mathbb{R}^d$ by

$$\|w\|_{\text{CAP}} = \|(\|w_{g_1}\|_{p_1}, \ldots, \|w_{g_n}\|_{p_n})\|_{p_0}^{p_0}. \tag{2.22}$$

In the case where $p_1 = \ldots = p_n$, the penalty computes the p_0-norm, raised to the power p_0, of the p_1-norms of the group components. When the groups form a partition, this is a straightforward variation of the group lasso, where the hyperparameters $p_0, \ldots, p_n$ can be tuned to the data.

A key use of the penalty is hierarchical selection when the underlying structure can be captured by a directed acyclic graph (DAG). In this setting, whenever a node corresponding to a non zero coordinate in the regression vector is activated, all predecessors of the node should also be activated. This behaviour can be encoded using the CAP penalty as whenever a subset of the groups have a nested structure, indicating a series of parent-child relationships in the graph, sparsity in the parents cascades down to the children. By way of example, the relationship w_1(parent) → w_2(child) could be encoded using groups $g_1 = \{1, 2\}$ and $g_2 = \{2\}$.

Letting $p_1 = p_2 = p > 1$ for simplicity, the corresponding components of the CAP penalty become

$$\|(w_1, w_2)\|_p^{p_0} + \|(w_2)\|_p^{p_0}.$$

The authors show that when the value of w_2 is non zero, the value of w_1 tends also to be non zero, as there is no additional penalty for this, that is whenever the child is activated, the parent is also activated.

2.4.7 Hierarchical models

A related approach to the CAP penalty for hierachical penalization was introduced by Yuan et al. (2009). Following on from the nonnegative garrote of Breiman (1995), who considered the square loss with constrained ℓ_1-norm, subject to nonnegative regression vector components, Yuan et al. (2009) add the constraint that for $G = \{g_1, \ldots, g_n\}$, for each g_i, the corresponding vector components w_{g_i} are nonnegative and nonincreasing. A similar method was proposed by Szafranski et al. (2007), where again an a priori set of groups determine a hierarchical structure. These methods do not explicitly define norms in the general case where $p_0 \neq 1$, and we do not elaborate on them further.

2.4.8 Overlapping group lasso

With the standard group lasso, the groups form a partition of $\mathbb{N}_d$. With the overlapping group lasso, this assumption is no longer enforced, allowing overlap between the groups, provided that their union $G = \{g_1, \ldots, g_n\}$ spans $\mathbb{N}_d$ (Jenatton et al., 2011b). The expression of the penalty is the same as (2.19). The latent group lasso can also be seen as a special case of the CAP in (2.22), whereby the principle norm is the ℓ_1-norm, that is $p_0 = 1$, and all others are set to the ℓ_2-norm (Zhao et al., 2009).

2.4.9 Ordered weighted ℓ_1-norm

The lasso has a single regularization parameter that regulates the threshhold below which components of the regression vector are set to zero, and this impacts all coordinates equally. The ordered weighted ℓ_1-norm was proposed to allow additional flexibility with respect to the individual components. Specifically, given $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$, the penalty is defined as

$$\|w\|_{\text{OWL}} = \sum_{i=1}^{d} \lambda_i |w|_i^{\downarrow},$$

where $|w|_i^{\downarrow}$ denotes the ith largest component of the vector in absolute value (Bogdan et al., 2013; Zeng and Figueiredo, 2014; Figueiredo and Nowak, 2016). While this requires d parameters, Bogdan et al. (2013) show that these can be chosen based on quantiles of a normal distribution in a manner which gives good statistical performance.

The penalties discussed so far have had explicit closed form definitions. A number of general mathematical frameworks have been proposed to generate families of norms, and we mention a few of these in the following sections. In particular, we highlight a number of norms which are defined using a variational formulation, that is, in order to compute the norm, it is necessary to solve an optimization problem, and this will be a common theme throughout later chapters of this thesis. In general, this approach leads to penalties capable of expressing more complex relationships between the variables. As we see in the following examples, a common approach in frameworks aimed at promoting sparse vectors involves constraining their support, or the support of component parts which decompose a vector.

2.4.10 Group lasso with overlap

The group lasso with overlap is a generalization of the group lasso. Given groups $G = \{g_1, \ldots, g_n\}$, whose union is $\mathbb{N}_d$ and which may overlap, the penalty is defined, for $w \in \mathbb{R}^d$, as

$$\|w\|_{\mathrm{GLO}} = \inf\left\{ \sum_{j=1}^{m} \|z_j\|_p \;\middle|\; z_j \in \mathbb{R}^d, \mathrm{supp}(z_j) \subset g_j, \sum_{j=1}^{m} z_j = w \right\}, \tag{2.23}$$

where $p \in [1, \infty]$ (Jacob et al., 2009b; Bach et al., 2012). As the groups may overlap, for a given vector w, the set of groups whose union includes $\mathrm{supp}(w)$ may not be unique, hence the definition involves a variational formulation which implicitly selects the most efficient subset of groups as measured by the sum of p-norms of the associated vectors. When the groups form a partition, the infimum is vacuous, and the penalty reduces to the standard group lasso. As with the CAP, this penalty can be motivated by a graph structure, whereby the nodes correspond to features, and the groups are determined by the connectivity of the graph, for instance corresponding to edges or cliques. In this case Jacob et al. (2009b) refer to the penalty as the graph lasso.

Note that the support recovery properties of the two group lasso approaches are complementary. With the standard group lasso, the complement of the support corresponding to components which are zero is always equal to the union of some subset of the groups. It follows that the group lasso of Yuan and Lin (2006) will never estimate a model whose support is a union of groups, whereas the group lasso with overlap does precisely this (Jacob et al., 2009b; Wainwright, 2014), and this was one of the motivating factors behind the introduction of the norm.

2.4.11 The k-support norm

The k-support norm was introduced by Argyriou et al. (2012) as an alternative regularizer to the ℓ_1-norm for sparse vector estimation problems. As outlined above, one motivation for the ℓ_1-norm stems from the fact that it is the convex relaxation of the card operator on the unit ℓ_∞-ball. Notwithstanding this fact, Argyriou et al. (2012) consider the convex hull of the set of k-sparse vectors with unit Euclidean norm and show that the k-support norm is a strictly tighter relaxation compared to the ℓ_1-norm. Its unit ball is the set C_k defined by

$$\operatorname{conv}\left\{w \in \mathbb{R}^d \;\middle|\; \operatorname{card}(w) \leq k, \|w\|_2 \leq 1\right\},$$

where $k \in \mathbb{N}_d$. From the definition of the unit ball, for $k = 1$ we recover the ℓ_1-norm, whereas for $k = d$ we recover the ℓ_2-norm. Used as a regularizer, for values of k in between these extremes, the norm has an intermediate effect, promoting sparsity in general, but less so than the ℓ_1-norm. The norm is a particular example of the group lasso with overlap (2.23), where the set of groups G is chosen to contain all groups with cardinality no greater than k. The k-support norm plays a central role throughout this thesis, and we will study it in further detail in the subsequent chapters.

2.4.12 Atomic norms

The group lasso with overlap, with the k-support as a particular example, can be expressed as an infimal convolution (see Rockafellar, 1970, p. 34). A general family of norms of this construction was studied by Chandrasekaran et al. (2012b), who consider norms of the type

$$\|w\|_{\mathcal{A}} = \inf\left\{\sum_{a \in \mathcal{A}} c_a \;\middle|\; w = \sum_{a \in \mathcal{A}} c_a a, c_a \geq 0\right\},$$

where $\mathcal{A}$ corresponds to the unit ball of the corresponding norm and is composed of elements which define the convex hull of $\mathcal{A}$. The formulation encompasses several families including the group lasso with overlap. The authors investigate the statistical recovery properties of the norms and describe how these depend on the geometry of the atomic set, as we highlighted in Section 2.3, see also Amelunxen et al. (2014) for a generalization of this idea. We also mention a related approach by Negrinho and Martins (2014), who apply group operators to a parameter vector to define the unit ball of a penalty as the convex hull of the images of the parameter. For appropriate settings this captures several of the norms reviewed in this chapter, and the authors provide optimization methods to address regularization problems with the penalties.

2.4.13 Submodular functions and polyhedral norms

As outlined above, one motivation for the use of sparsity inducing penalties deriving from the ℓ_1-norm stems from the fact that the convex relaxation of the card operator on the unit ℓ_∞-norm ball is the ℓ_1-norm. An alternate approach is taken by Bach (2010, 2013), who start from a submodular set-function $w \mapsto F(\operatorname{supp}(w))$ and derive a norm via its convex envelope on the unit ℓ_∞-norm ball. Each such norm is polyhedral, that is, its unit ball has finite extreme points and is defined by the intersection of half spaces.

A real valued function F defined on the power set V of $\mathbb{N}_d$ is nondecreasing submodular if for all $A, B \subset V$, we have

$$F(A) + F(B) \geq F(A \cup B) + F(A \cap B), \text{ and}$$
$$A \subset B \Rightarrow F(A) \leq F(B),$$

and we assume $F(\emptyset) = 0$. For each such function F, its Lovasz extension f is a convex function which extends the former from the discrete domain V to the continuous domain $\mathbb{R}^d$. Specifically, this is computed at a vector $w \in \mathbb{R}^d_+$ by ordering the components in decreasing order $w_{j_1} \geq \ldots \geq w_{j_d} \geq 0$, and by defining

$$f(w) = \sum_{k=1}^{d} w_{j_k} [F(\{j_1, \ldots, j_k\}) - F(\{j_1, \ldots, j_{k-1}\})].$$

The extension property of function f uses the fact that for $\delta \in \{0,1\}^d$, $f(\delta) = F(\mathrm{supp}(\delta))$, where we can consider δ an indicator vector. Finally, if the function F is strictly positive for all singletons, then $\Omega : w \mapsto f(|w|)$ defines a norm, whose unit ball is polyhedral. The procedure encompasses various existing norms. For example, choosing f to be the cardinality function induces the ℓ_1-norm. Similarly, if $G = \{g_1, \ldots, g_n\}$ is a partition of $\mathbb{N}_d$, the set function $A \mapsto F(A)$, which is equal to the number of groups $B_{g_1}, \ldots, B_{g_n}$ with non empty intersection with A, induces the CAP penalty (Zhao et al., 2009).

2.4.14 Structured sparsity Λ-norms

We now discuss the class of norms introduced by Micchelli et al. (2010). The class is defined by a prescribed conic subset Λ of the positive orthant, and we refer to them as Λ-norms throughout this thesis. Specifically, for the cone $\Lambda \subset \mathbb{R}^d_{++}$, the norm is defined, for $w \in \mathbb{R}^d$ by

$$\|w\|_\Lambda = \inf_{\lambda \in \Lambda} \left\{ \frac{1}{2} \sum_{i=1}^{d} \left(\frac{w_i^2}{\lambda_i} + \lambda_i \right) \right\}. \tag{2.24}$$

The authors show that this is a norm, and for fixed $\lambda \in \mathbb{R}^d_{++}$ the objective in (2.24) provides a smooth approximation to $\|w\|_1$ from above, which is exact at $w = \pm\lambda$. Furthermore, the corresponding dual norm is given, for $u \in \mathbb{R}^d$, by

$$\|u\|_{\Lambda,*} = \sup_{\lambda \in \Lambda} \left\{ \sqrt{\frac{\sum_{i=1}^{d} \lambda_i u_i^2}{\sum_{i=1}^{d} \lambda_i}} \right\}. \tag{2.25}$$

The construction of the norm and its dual via the parameter set Λ allows for different models of sparsity patterns to be encoded, including more complex relationships than the group lasso with overlap discussed earlier. One example is the so-called wedge penalty (Micchelli et al., 2010). We define the wedge parameter set as the positive monotone cone

$$\Lambda_W = \{\lambda \in \mathbb{R}^d_{++} \mid \lambda_1 \geq \ldots \geq \lambda_d\}, \tag{2.26}$$

that is the subset of the positive orthant which corresponds to vectors with nonincreasing components. The Λ-norm induced by the set Λ_W, when used as a regularizer for regression with the square loss, promotes solutions whose components are ordered in the same fashion. In Chapter 3 we will introduce and discuss in detail a closely related family of norms which subsumes these.

2.4.15 Convex relaxation of set functions

A recent work by Obozinski and Bach (2016) follows on from Bach (2010), and considers norms which are defined as the convex envelope of a set-valued function F of the support of a vector, plus an ℓ_p-norm, raised to a power. Specifically they consider the penalty $\mu F(\mathrm{supp}(w)) + \nu\|w\|_p^p$, for $\mu, \nu > 0$, where F is defined on the power set V of $\{1, \ldots, d\}$ such that $F(\emptyset) = 0$, and $F(A) \leq F(V)$ for any $A \subset V$. They define the penalty $h(w) = F(\mathrm{supp}(w))^{\frac{1}{q}}\|w\|_p$, where q satisfies $\frac{1}{p} + \frac{1}{q} = 1$. They show that the convex envelope of h is a multiple of a norm Ω_p, whose dual norm is defined by $\Omega_p^*(w) = \max_{A \subset V, A \neq \emptyset} \frac{\|w_A\|_q}{F(A)^{\frac{1}{q}}}$. For appropriate choices of F and p, Ω_p captures several norms discussed above, in particular subsuming the ℓ_p-norms, the k-support norm, the ordered weighted ℓ_1-norm and the wedge penalty of Micchelli et al. (2010). Furthermore, the norm Ω_p can be written as a variational expression involving the Lovasz extension, see Obozinski and Bach (2016) for details.

2.5 Matrix Penalties

In this section we discuss regularizers which have been proposed for matrices with sparse structure. Due to the additional structure in the space of matrices, there are different types of sparsity in the context of matrix learning problems. The first is sparsity in the elements. The zeros can be distributed throughout the matrix entries, or the matrix can have one or several zero rows or columns.

2.5.1 Sparsity in elements

Note that the matrix ℓ_1-norm, defined by $\|X\|_1 = \sum_{i,j}|X_{ij}|$, is equivalent to the vector ℓ_1-norm of the vectorized matrix $\mathrm{vec}(X)$. Consequently the norm can be used to encourage matrices with zero elements. Furthermore, for the Frobenius norm we have $\|X\|_{\mathrm{fro}} = \sqrt{\sum_{i,j} X_{ij}^2} = \|\mathrm{vec}(X)\|_2$, hence the norm acts to minimize outlier values, as with ridge regression in the vector case. In order to promote zero columns, the $\ell_{1,2}$ norm has been studied, see e.g. Yuan and Lin (2006). Similar to the case of the group lasso for vectors, the sum of ℓ_2 norms of columns promotes columns with zero Euclidean norm, which by positivity implies that the entire column vector is zero. For rows, the same method can be applied to the transpose of the matrix.

2.5.2 Common support sets

Norms which address the columns of a matrix have been used in the setting of multiple regression learning and multitask learning, which we expand on below. In particular, the $\ell_{1,\infty}$-norm is a special case of the CAP, (2.22), where the norms $\|\cdot\|_{p_1}, \ldots, \|\cdot\|_{p_n}$ are all set to the ℓ_∞-norm, and this has been studied in the case where the regression vectors share a common support set by Negahban and Wainwright (2011); Turlach et al. (2005); Tropp et al. (2006).

2.5.3 Sparsity in structure

In contrast to the vector case, sparsity in the elements does not necessarily prevent a matrix from being complex in some sense. Comparing a sparse, full rank, diagonal matrix to the dense, rank one, matrix whose entries are all identical, it is clear that presence or lack of zeros does not completely define the complexity of a matrix. With this in mind, the second type of sparsity for matrices relates to the inherent structure of the matrix. A large body of research has developed around factorization models to determine the characteristics of a matrix (Bach et al., 2008). With this approach, a matrix $W \in \mathbb{R}^{d \times m}$ is assumed to have factorization $W = UV^\top$, with $U \in \mathbb{R}^{d \times k}$, $V \in \mathbb{R}^{m \times k}$. This model captures a number of problem instances including clustering, dictionary learning, and non negative matrix factorization (Bach et al., 2008). In this thesis we focus on two matrix learning frameworks, low rank matrix learning and multitask learning. We also mention that dual approaches which promote element sparsity and low rankedness exist, see e.g. Savalle (2014).

2.5.3.1 Low rank matrix learning

In recent years there has been a great deal of interest in the problem of learning a low rank matrix from a set of linear measurements. A widely studied and successful instance of this problem arises in the context of matrix completion or collaborative filtering, in which we want to recover a low rank (or approximately low rank) matrix from a small sample of its entries, see e.g. Srebro et al. (2005); Abernethy et al. (2009) and references therein. In the factorization framework, this corresponds to taking $k \ll \min(d,m)$. One prominent method of solving this problem is trace norm regularization: we look for a matrix which closely fits the observed entries and has a small trace norm (sum of singular values) (Jaggi and Sulovsky, 2010; Toh and Yun, 2011; Mazumder et al., 2010).

2.5.3.2 Multitask learning

Multitask learning is framework for solving multiple related regression or classification problems simultaneously. Each problem, or task, is represented by a single weight vector, and these are stacked as the columns of a matrix W which is to be learned. The goal behind the method is to make use of commonalities between the tasks in order to improve learning relative to solving the tasks independently (Evgeniou et al., 2005; Argyriou et al., 2007a, 2008; Jacob et al., 2009a; Cavallanti et al., 2010). In particular, in instances where available training data for each task are limited, aggregating the data and solving the tasks together has been shown to improve learning (Maurer and Pontil, 2012). A general approach to multitask learning is using spectral regularizers to encourage low rank solutions (Evgeniou et al., 2005; Argyriou et al., 2008; Jacob et al., 2009a). When the tasks are clustered in several groups we refer to the problem as clustered multitask learning, which we discuss in Chapter 4.

2.5.4 Spectral regularizers

As mentioned above, the vector of singular values of a low rank matrix is sparse. One popular class of regularizers is given by spectral regularizers, which are norms which are functions of the singular values (Argyriou et al., 2010). When these norms are chosen to promote sparsity in the spectrum, the resulting matrix has low rank, and such regularizers have been applied to a range of problems including multiple regression, matrix completion, and multitask learning, see e.g. Evgeniou et al. (2005); Argyriou et al. (2008); Jacob et al. (2009a) and references therein.

Some intuition for this is given by the fact that given a matrix W, the closest rank k matrix in Frobenius norm is characterized by taking W and setting to zero all but the largest k singular values.

Lemma 49 (Best rank k matrix approximation). *Let $X \in \mathbb{R}^{d\times m}$ have singular value decomposition $X = U\operatorname{diag}(\sigma(X))V^\top$ and let $k \leq q = \min(d,m)$. Then the closest rank k matrix to X in Frobenius norm is given by the matrix $Z = U\operatorname{diag}(\sigma(Z))V^\top$, where $\sigma(Z) = (\sigma_1(X),\ldots,\sigma_k(X),0,\ldots,0)$.*

Proof. See Appendix A. □

Spectral norms have the property that they are invariant to left and right multiplication by orthogonal matrices, which is referred to as orthogonally invariant, that is for $W \in \mathbb{R}^{d\times m}$,

$$\|W\| = \|AWB^\top\| = \|\Sigma_W\|, \tag{2.27}$$

where $A \in \mathbf{O}^d$, $B \in \mathbf{O}^m$ are orthogonal, and Σ_W is the matrix of singular values of W. As discussed in Section 2.2.2, recall that a classical result by von Neumann states that a matrix norm is orthogonally invariant if and only if it is of the form $\|W\| = g(\sigma(W))$, where $\sigma(W)$ is the vector formed by the singular values of W, and g is a symmetric gauge function.

2.5.5 Frobenius norm

The standard Euclidean norm for matrices is the Frobenius norm, that is the square root of the sum of the squared entries, $\|W\|_{\text{fro}} = \sqrt{\sum_{ij} W_{ij}^2}$. Note that the squared Frobenius norm can equally be considered as the sum of the squared ℓ_2-norms of the columns or the rows of the matrix. Consequently, as with the Euclidean norm in the vector case, when used a regularizer, the Frobenius norm encourages the entries of the matrix to be similar in magnitude. Moreover, for $W = [w_1, \ldots, w_m] \in \mathbb{R}^{d \times m}$ we have

$$\|W\|_{\text{fro}}^2 = \sum_{j=1}^{m} \sum_{i=1}^{d} W_{ij}^2 = \sum_{j=1}^{m} \|w_j\|_2^2 = \sum_{j=1}^{m} w_j^\top w_j = \operatorname{tr}(W^\top W),$$

and letting W have singular value decomposition $W = U \Sigma_W V^\top$, we have

$$\operatorname{tr}(W^\top W) = \operatorname{tr}(V \Sigma_W U^\top U \Sigma_W V^\top) = \operatorname{tr}(\Sigma_W^\top \Sigma_W) = \|\sigma(W)\|_2^2,$$

using the cyclic property of the *trace* operator, hence $\|W\|_{\text{fro}} = \sqrt{\operatorname{tr}(W^\top W)} = \|\sigma(W)\|_2$, and we see that the Frobenius norm is the orthogonally invariant spectral norm induced by the ℓ_2-norm.

2.5.6 Trace norm

The baseline regularizer for matrices is the orthogonally invariant norm induced by the ℓ_1-norm, the trace norm (or nuclear norm)

$$\|W\|_{\text{tr}} = \|\sigma(W)\|_1 = \sum_{i=1}^{r} \sigma(W)_i, \quad W \in \mathbb{R}^{d \times m},$$

where $r = \min(d, m)$. Used as a regularizer in matrix learning problems, the norm promotes low rank solutions (Argyriou et al., 2007b; Mazumder et al., 2010; Srebro et al., 2005; Toh and Yun, 2011). It has also been used in multitask learning problems, which we return to in Chapter 4. The use of the ℓ_1-norm as a convex approximation of the cardinality operator for vectors naturally carries over to the trace norm as a convex approximation of the rank operator for matrices (Jojic et al., 2011).

The trace norm can be computed directly via the singular value decomposition. It can also be expressed in a variational formulation

$$\|W\|_{\text{tr}} = \frac{1}{2} \inf_{S \succeq 0} \operatorname{tr}(W^\top S^{-1} W + S),$$

see e.g. Argyriou et al. (2008) and Bach et al. (2012). In subsequent chapters we focus in particular on norms which have variational definitions. In particular, in Chapter 4, we use a related formulation when we discuss the cluster norm and the spectral k-support norm.

2.5.7 Schatten p-norms, Ky-Fan k-norms

The Frobenius and trace norms are both examples of the Schatten p-norms. These are given, for $W \in \mathbb{R}^{d \times m}$ and $p \in [1, \infty]$ by

$$\|W\|_p = \|\sigma(W)\|_p = \left(\sum_{i=1}^{r} \sigma_i(W)^p \right)^{\frac{1}{p}},$$

where $r = \min(d, m)$, that is they are the orthogonally invariant matrix norms induced by the vector ℓ_p-norms. We return to these in Chapter 4.

2.5.8 Cluster norm

The cluster norm was introduced by Jacob et al. (2009a) as a regularizer for multitask learning problems where the regression vectors for the tasks are understood to exhibit clustered behavior. The norm is defined by the following variational formulation. Letting $\mathcal{S}$ be the set of $d \times d$ matrices whose singular values lie in the interval $[a, b]$, and whose trace is upper bounded by c, for specific values of a, b, and c, the cluster norm is defined for $n \times d$ matrices W as

$$\|W\|_{\mathrm{cl}} = \sqrt{\inf_{\Sigma \in \mathcal{S}} \operatorname{tr}(W \Sigma^{-1} W^\top)}.$$

Used as a regularizer in multitask learning problems, where the penalty is computed by first centering the columns of W, the cluster norm promotes clustering in the regression vectors. We discuss the cluster norm further in Chapter 4, where we will show its connection to the k-support norm of Argyriou et al. (2012).

2.5.9 Variational Gram functions

A recent work by Jalali et al. (2016) introduced a class of functions which promote pairwise relations, such as orthogonality, among the columns of a matrix. The penalty $\Omega_{\mathcal{M}}$, which the authors name a *variational Gram function* (VGF), is defined, for $X \in \mathbb{R}^{d \times m}$, as

$$\Omega_{\mathcal{M}}(X) = \max_{M \in \mathcal{M}} \operatorname{tr}(X M X^\top), \tag{2.28}$$

where $\mathcal{M}$ is a subset of $\mathbf{S}^m$. The authors give conditions under which $\Omega_{\mathcal{M}}$ is convex, and define the penalty $\|\cdot\|_{\mathcal{M}} = \sqrt{\Omega_{\mathcal{M}}(\cdot)}$. The formulation is general and captures a number of well known norms. Given any vector norm $\|\cdot\|$ on $\mathbb{R}^d$, considering the VGF defined by the set $\mathcal{M} = \left\{ uu^\top \mid \|u\|_* \leq 1 \right\}$, using $\operatorname{tr}(x^\top u u^\top x) = \langle u \mid x \rangle^2$, we see that we recover the original vector norm in the induced penalty $\|\cdot\|_{\mathcal{M}}$. Furthermore, the k-support norm corresponds to the choice of $\mathcal{M} = \left\{ \operatorname{diag}\theta \mid 0 < \theta \leq 1, \|\theta\|_1 \leq k \right\}$, a result that we prove and subsequently use in Chapter 3. If the set $\mathcal{M}$ is closed under left and right multiplication by orthogonal matrices, the VGF is a spectral function. In Chapter 4 we introduce a matrix k-support norm and its Moreau envelope the *box norm*, and we show its relation to the cluster norm of Jacob et al. (2009a). The box norm then corresponds to a VGF for an appropriate choice of $\mathcal{M}$, and we will return to this characterization.

2.6 Discussion

In this chapter we introduced our learning framework. In each case, the problem is framed as a regularized or constrained optimization problem, and we reviewed core concepts from convex analysis and optimization that can be applied to the respective problems. We concluded that problems involving convex penalties and convex loss functions can be efficiently tackled using tools from convex analysis and optimization, and this motivates the use of norms as penalties. We surveyed a number of norms which have been proposed in the literature as structured sparsity penalties, starting from the ℓ_1-norm, and progressing to more sophisticated norms. We noted that the increased complexity of the norms allows us to more finely tune the learning process, generally improving the performance of the penalties both theoretically and empirically. In particular, in recent years, a range of complex penalties have been proposed that themselves are defined as an optimization problem. While this brings additional computational complexities, the variational formulations can be used to derive numerical methods, and to impose more sophisticated properties than the simple penalties, leading to better performance.

In the subsequent chapters of this thesis we study a number of norms that have variational formulations, and which can be used in sparse vector, matrix and tensor problems. In particular, we propose a number of norms that are defined as an infimal convolution, we study their properties and their relationship to existing norms, and we demonstrate that the additional complexity can lead to improved performance on a number of real world datasets. Finally, we present a general interpolation framework that captures many of the norms studied in this thesis.

Chapter 3

The Θ-norm for structured sparsity

In Chapter 2 we reviewed several norms that have been studied as penalty functions in structured sparsity learning problems. In particular, penalties have recently been introduced that are defined as variational problems, and the added complexity has led to improved learning. In this chapter we proceed in this vein and we continue the study of a family of norms that are obtained by taking the infimum of a class of quadratic functions. These norms can be used as a regularizer in linear regression learning problems, where the parameter set can be tailored to assumptions on the structure of the underlying regression model. Our work builds upon a recent line of papers that in addition to families of quadratics, also considered related infimal convolution problems, see e.g. Jacob et al. (2009b); Bach et al. (2012); Maurer and Pontil (2012); Micchelli and Pontil (2005); Obozinski and Bach (2012) and references therein. Related variational formulations for the lasso have also been discussed in Grandvalet (1998) and further studied in Szafranski et al. (2007).

This family of norms is rich and encompasses a number of regularizers outlined in Chapter 2, such as the ℓ_p norms, the norm of Micchelli et al. (2013) and the group lasso with overlap (Jacob et al., 2009b) which includes the k-support norm introduced by Argyriou et al. (2012), and which we study in more detail in this chapter. In particular, we show that the norm is a special case of the box-norm, in which the parameter set involves box constraints and a linear constraint. We show that it can be generated as a perturbation of the k-support norm introduced by Argyriou et al. (2012) for sparse vector estimation, which hence can be seen as a special case of the box-norm. Furthermore, our variational framework allows us to study efficient algorithms to compute the norms and the proximity operator of the square of the norms.

This chapter is organized as follows. In Section 3.1, we review a general class of norms and characterize their unit ball. In Section 3.2, we specialize these norms to the box-norm, which we show is a perturbation of the k-support norm. We study the properties of the norms and we describe the geometry of the unit balls. In Section 3.3, we compute the box-norm and we provide an efficient method to compute the proximity operator of the squared norm. Finally in Section 3.4 we discuss a generalization of the Θ-norm which we use in Chapter 5 to generalize the k-support norm.

3.1 The Θ-Norm

In this section we review a family of norms parameterized by a set Θ that we call the Θ-norms. They are closely related to the norms considered in Micchelli et al. (2010, 2013). Similar norms are also discussed in Bach et al. (2012, Sect. 1.4.2) where they are called H-norms. We first recall the definition of the norm.

Definition 50. *Let Θ be a convex bounded subset of the open positive orthant. For $w \in \mathbb{R}^d$ the Θ-norm is defined as*

$$\|w\|_\Theta = \sqrt{\inf_{\theta \in \Theta} \sum_{i=1}^{d} \frac{w_i^2}{\theta_i}}. \tag{3.1}$$

Note that the function $(w,\theta) \mapsto \sum_{i=1}^{d} \frac{w_i^2}{\theta_i}$ is strictly convex on $\mathbb{R}^d \times \mathbb{R}^d_{++}$, hence every minimizing sequence converges to the same point. The infimum, however, is not attained in general because a minimizing sequence may converge to a point on the boundary of Θ. For instance, if $\Theta = \big\{\theta \in \mathbb{R}^d_{++} \;\big|\; \sum_{i=1}^{d} \theta_i \leq 1\big\}$, then $\|w\|_\Theta = \|w\|_1$ and the minimizing sequence converges to the point $\big(\frac{|w_1|}{\|w\|_1}, \ldots, \frac{|w_d|}{\|w\|_1}\big)$, which belongs to Θ only if all the components of w are different from zero.

Proposition 51. *The Θ-norm is well defined and the dual norm is given, for $u \in \mathbb{R}^d$, by*

$$\|u\|_{\Theta,*} = \sqrt{\sup_{\theta \in \Theta} \sum_{i=1}^{d} \theta_i u_i^2}. \tag{3.2}$$

Proof. Consider the expression for the dual norm. The function $\|\cdot\|_{\Theta,*}$ is a norm since it is a supremum of norms. Recall that the Fenchel conjugate h^* of a function $h : \mathbb{R}^d \to \mathbb{R}$ is defined for every $u \in \mathbb{R}^d$ as $h^*(u) = \sup\big\{\langle u \mid w\rangle - h(w) \;\big|\; w \in \mathbb{R}^d\big\}$. It is a standard result from convex analysis that for any norm $\|\cdot\|$, the Fenchel conjugate of the function $h := \frac{1}{2}\|\cdot\|^2$ satisfies $h^* = \frac{1}{2}\|\cdot\|_*^2$, where $\|\cdot\|_*$ is the corresponding dual norm, see e.g. Example 38 in Chapter 2 and Lewis (1995). By the same result, for any norm the biconjugate is equal to the norm, that is $(\|\cdot\|^*)^* = \|\cdot\|$. Applying this to the dual norm we have, for every $w \in \mathbb{R}^d$, that

$$h(w) = \sup_{u \in \mathbb{R}^d} \{\langle w \mid u\rangle - h^*(u)\} = \sup_{u \in \mathbb{R}^d} \inf_{\theta \in \Theta} \left\{ \sum_{i=1}^{d} \left(w_i u_i - \frac{1}{2}\theta_i u_i^2 \right) \right\}.$$

This is a minimax problem in the sense of von Neumann (see e.g. Prop. 2.6.3 in Bertsekas et al., 2003), and we can exchange the order of the $\inf$ and the $\sup$, and solve the latter (which is in fact a maximum) componentwise. The gradient with respect to u_i is zero for $u_i = \frac{w_i}{\theta_i}$, and substituting this into the objective function we obtain $h(w) = \frac{1}{2}\|w\|_{\Theta}^2$. It follows that the expression in (3.1) defines a norm, and its dual norm is defined by (3.2), as required. □

3.1.1 Examples

The Θ-norm (3.1) encompasses a number of well known norms. For example, they include the ℓ_p-norms for $p \in [1,\infty]$ as follows, see also Micchelli and Pontil (2005, Lemma 26), and Aflalo et al. (2011) who considered the case of $p > 2$. Recall that for $p \in [1,\infty)$ the ℓ_p norm is defined, for every $w \in \mathbb{R}^d$, as $\|w\|_p = \left(\sum_{i=1}^d |w_i|^p\right)^{\frac{1}{p}}$, if $p \in [1,\infty)$ and $\|w\|_\infty = \max_{i=1}^d |w_i|$.

Lemma 52. *For $p \in [1,\infty]$ we have*

$$\|w\|_p = \begin{cases} \|w\|_\Theta, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \sum_{i=1}^d \theta_i^{\frac{p}{2-p}} \leq 1\right\}, \quad \textit{for } p \in [1,2), \\ \|w\|_\Theta = \|w\|_{\Theta,*}, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; 0 < \theta_i \leq 1\right\}, \quad \textit{for } p = 2, \\ \|w\|_{\Theta,*}, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \sum_{i=1}^d \theta_i^{\frac{p}{p-2}} \leq 1\right\}, \quad \textit{for } p \in (2,\infty]. \end{cases}$$

Note that the case $p = 2$ follows from the cases $p \in [1,2)$, respectively $p \in (2,\infty]$, by taking the limit as p tends to 2 from below, respectively from above.

Proof. See Appendix B. □

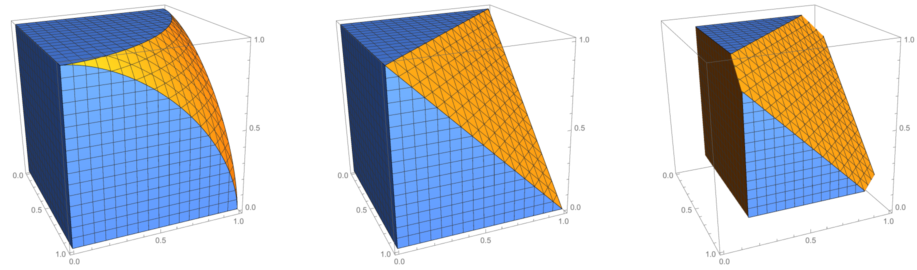
Figure 3.1 (left) illustrates the Θ parameter set for the ℓ_p-norm ($p \in \{\frac{5}{4}, 5\}$) in $\mathbb{R}^3$.

Other norms which belong to the family (3.1) are the so called Λ-norms for structured sparsity of Micchelli et al. (2013) that we recalled in Section 2.4.14. A specific example described therein is the wedge penalty, which corresponds to choosing $\Lambda = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \theta_1 \geq \ldots \geq \theta_d\right\}$. The equivalence is given by choosing $\Theta = \left\{\theta \in \Lambda \;\middle|\; \sum_{i=1}^d \theta_i \leq 1\right\}$, where $\Lambda \subset \mathbb{R}^d_{++}$ is a convex cone, as we now show. Letting Λ be a conic subset of $\mathbb{R}^d_{++}$, recall that the primal and dual norms are defined by

$$\|w\|_\Lambda = \inf_{\lambda \in \Lambda} \left\{ \frac{1}{2} \sum_{i=1}^d \left(\frac{w^2}{\lambda_i} + \lambda_i \right) \right\} \tag{3.3}$$

$$\|u\|_{\Lambda,*} = \sup_{\lambda \in \Lambda} \left\{ \sqrt{\frac{\sum_{i=1}^d \lambda_i w^2}{\sum_{i=1}^d \lambda_i}} \right\}. \tag{3.4}$$

Figure 3.1: Θ sets for ℓ_p-norm ($p \in \{\frac{5}{4}, 5\}$), k-support norm ($k = 2$) and box-norm ($a = 0.15$, $b = 1$) in $\mathbb{R}^3$.



Lemma 53. *Let Λ be a conic subset of $\mathbb{R}^d_{++}$. Define $\Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \theta = \frac{\lambda}{\|\lambda\|_1}, \lambda \in \Lambda\right\}$. Then*

$$\|w\|_\Theta = \|w\|_\Lambda,$$

$$\|w\|_{\Theta,*} = \|w\|_{\Lambda,*}.$$

Proof. First note that Θ is a one-dimensional manifold given by the intersection between the cone Λ and the unit simplex (ℓ_1 unit ball) in the strictly positive orthant. It follows that it is convex and bounded, hence the Θ-norms are well defined. Considering the dual norm we have

$$\begin{aligned}\|w\|^2_{\Theta,*} &= \sup_{\theta\in\Theta} \sum_{i=1}^d \frac{w_i^2}{\theta_i} = \sup_{\lambda\in\Lambda} \sum_{i=1}^d \left(\frac{\lambda_i}{\sum_{j=1}^d \lambda_j}\right) w_i^2 = \sup_{\lambda\in\Lambda} \frac{\sum_{i=1}^d \lambda_i w_i^2}{\sum_{j=1}^d \lambda_j}\\ &= \|w\|_{\Lambda,*}.\end{aligned}$$

As the dual norms are equivalent, it follows that the primal norms are also equivalent. □

The Θ-norm has a direct dependence on the scaling of the set Θ.

Lemma 54 (Scaling)**.** *Let Θ be a convex bounded subset of the open positive orthant, $c > 0$, and define $\tilde{\Theta} = c^2\Theta$. Then the set $\tilde{\Theta}$ defines a primal and dual norm, and there hold $\|\cdot\|_{\tilde{\Theta}} = \frac{1}{c}\|\cdot\|_\Theta$ and $\|\cdot\|_{\tilde{\Theta},*} = c\|\cdot\|_{\Theta,*}$.*

Proof. As $c > 0$, the set $\tilde{\Theta}$ is a convex bounded subset of the open positive orthant, hence the primal and dual norms are well defined. The scaling result follows by direct computation, or by noting that with respect to each $\frac{1}{\theta_i}$ (respectively θ_i), the primal (respectively dual) norm is a homogeneous function of order $\frac{1}{2}$. □

For the remainder of this section we consider the case where Θ is a polyhedron. As we show below, this setting applies to a number of norms of practical interest, including the group lasso with overlap, the wedge norm mentioned above and, in particular the k-support norm, which forms a central theme throughout this thesis. We first characterize the unit ball of the norm, which allows us to identify the norm as promised. To describe our observation, for every vector $\gamma \in \mathbb{R}^d_+$, we define the seminorm

$$\|w\|_\gamma = \sqrt{\sum_{i:\gamma_i>0} \frac{w_i^2}{\gamma_i}}.$$

Proposition 55. *Let $\gamma^1, \ldots, \gamma^m \in \mathbb{R}^d_+$ such that $\sum_{\ell=1}^m \gamma^\ell \in \mathbb{R}^d_{++}$ and let Θ be defined by $\Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \theta = \sum_{\ell=1}^m \lambda_\ell \gamma^\ell,\ \lambda \in \Delta^{m-1}\right\}$. Then we have, for every $w \in \mathbb{R}^d$, that*

$$\|w\|_\Theta = \inf\left\{\sum_{\ell=1}^m \|v_\ell\|_{\gamma^\ell} \;\middle|\; v_\ell \in \mathbb{R}^d,\ \operatorname{supp}(v_\ell) \subset \operatorname{supp}(\gamma^\ell),\ \ell \in \mathbb{N}_m,\ \sum_{\ell=1}^m v_\ell = w\right\}. \tag{3.5}$$

Moreover, the unit ball of the norm is given by the convex hull of the set

$$\bigcup_{\ell=1}^m \left\{w \in \mathbb{R}^d \;\middle|\; \operatorname{supp}(w) \subset \operatorname{supp}(\gamma^\ell), \|w\|_{\gamma^\ell} \le 1\right\}. \tag{3.6}$$

Proof. The proof of this result is presented in Appendix B. It is based on observing that the Minkowski functional (see e.g. Rudin, 1991) of the convex hull of the set (3.6) is a norm and it is given by the right hand side of equation (3.5); we then prove that this norm coincides with $\|\cdot\|_\Theta$ by noting that both norms share the same dual norm. □

The proof of Proposition 55 reveals that the unit ball of the dual norm of $\|\cdot\|_\Theta$ is given by an intersection of ellipsoids in $\mathbb{R}^d$. Indeed, we find that

$$\begin{aligned}\left\{u \in \mathbb{R}^d \mid \|u\|_{\Theta,*} \le 1\right\} &= \left\{u \in \mathbb{R}^d \;\middle|\; \max_{\ell=1}^m \sum_{i=1}^d \gamma_i^\ell u_i^2 \le 1\right\}\\ &= \left\{u \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^d \gamma_i^\ell u_i^2 \le 1, \forall \ell \in \mathbb{N}_m\right\}\\ &= \bigcap_{\ell \in \mathbb{N}_m} \left\{u \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^d \gamma_i^\ell u_i^2 \le 1\right\},\end{aligned}$$

see Equation (B.8) in Appendix B. Notice that for each $\ell \in \mathbb{N}_m$, the set $\left\{u \in \mathbb{R}^d \;\middle|\; \sum_{i=1}^d \gamma_i^\ell u_i^2 \le 1\right\}$ defines a (possibly degenerate) ellipsoid in $\mathbb{R}^d$, where the i-th semi-principal axis has length $1/\sqrt{\gamma_i^\ell}$ (which is infinite if $\gamma_i^\ell = 0$) and the unit ball of the dual Θ-norm is given by the intersection of m such ellipsoids. We return to this in Section 3.2.2.

To illustrate an application of the proposition, we specialize it to the group Lasso with overlap (Jacob et al., 2009b), of which the k-support norm is an example.

Corollary 56. *If $\mathcal{G}$ is a collection of subsets of $\mathbb{N}_d$ such that $\bigcup_{g\in\mathcal{G}} g = \mathbb{N}_d$ and Θ is the interior* *of the set* $\operatorname{co}\{1_g \mid g \in \mathcal{G}\}$*, then we have, for every* $w \in \mathbb{R}^d$*, that*

$$\|w\|_\Theta = \inf\left\{ \sum_{g\in\mathcal{G}} \|v_g\|_2 \;\middle|\; v_g \in \mathbb{R}^d,\ \operatorname{supp}(v_g) \subset g,\ \sum_{g\in\mathcal{G}} v_g = w \right\}. \tag{3.7}$$

Moreover, the unit ball of the norm is given by the convex hull of the set

$$\bigcup_{g\in\mathcal{G}} \left\{ w \in \mathbb{R}^d \;\middle|\; \operatorname{supp}(w) \subset g, \|w\|_2 \le 1 \right\}. \tag{3.8}$$

The utility of the result is that it links seemingly different norms such as the group Lasso with overlap and the Θ-norms, which provide a more compact representation, involving only d additional variables. This formulation is especially useful whenever the optimization problem (3.1) can be solved in closed form. One such example is provided by the wedge penalty described above.

As highlighted earlier we have the following characterization of the unit ball of the k-support norm.

Corollary 57. *The unit ball of the k-support norm is equal to the convex hull of the set* $\{w \in \mathbb{R}^d \mid \operatorname{card}(w) \le k, \|w\|_2 \le 1\}$.

Proof. The result follows directly by Corollary 56 for $\mathcal{G} = \mathcal{G}_k$ observing that in this case $\bigcup_{g\in\mathcal{G}_k} \{w \in \mathbb{R}^d \mid \operatorname{supp}(w) \subset g, \|w\|_2 \le 1\} = \{w \in \mathbb{R}^d \mid \operatorname{card}(w) \le k, \|w\|_2 \le 1\}$. □

3.2 The Box-Norm and the k-Support Norm

The Θ-norm formulation allows us to characterize the connection between the k-support norm and a regularizer which we introduce below, and the symmetries of the parameter set Θ will lead to matrix regularizers which we discuss in Chapter 4. In particular, we specialize our analysis to the case that

$$\Theta = \left\{ \theta \in \mathbb{R}^d \;\middle|\; a \le \theta_i \le b, \sum_{i=1}^d \theta_i \le c \right\} \tag{3.9}$$

where $0 < a \le b$ and $c \in [ad, bd]$. We call the norm defined by (3.1) the *box-norm* and we denote it by $\|\cdot\|_{\text{box}}$.

The structure of set Θ for the box-norm will be fundamental in computing the norm and deriving the proximity operator in Section 3.3. Furthermore, we note that the constraints are invariant with respect to permutations of the components of Θ and, as we shall see in Section 4.1, this property is key to extending the norm to matrices. Finally, while a restriction of the general family, the box-norm nevertheless encompasses a number of norms including the ℓ_1 and ℓ_2 norms, as well as the k-support norm, which we now recall.

For every $k \in \mathbb{N}_d$, the k-support norm $\|\cdot\|_{(k)}$ is defined as the norm whose unit ball is the convex hull of the set $S_k^{(2)}$ of vectors of cardinality at most k and ℓ_2-norm no greater than one. The norm was introduced by Argyriou et al. (2012) with the goal of learning a sparse vector which has small ℓ_2-norm, and they further show that the k-support norm is a tighter relaxation of $S_k^{(2)}$ than the elastic net. They further show that the k-support norm can be written as the infimal convolution (see Rockafellar, 1970, p. 34)

$$\|w\|_{(k)} = \inf\left\{ \sum_{g\in\mathcal{G}_k} \|v_g\|_2 \;\middle|\; v_g \in \mathbb{R}^d,\ \mathrm{supp}(v_g) \subset g,\ \sum_{g\in\mathcal{G}_k} v_g = w \right\}, \quad w \in \mathbb{R}^d, \tag{3.10}$$

where $\mathcal{G}_k$ is the collection of all subsets of $\mathbb{N}_d$ containing at most k elements. The k-support norm is a special case of the group lasso with overlap (Jacob et al., 2009b), where the cardinality of the support sets is at most k. When used as a regularizer, the norm encourages vectors w to be a sum of a limited number of vectors with small support. Note that while definition (3.10) involves a combinatorial number of variables, Argyriou et al. (2012) observed that the norm can be computed in $\mathcal{O}(d \log d)$, a point we return to in Section 3.3.

Comparing equation (3.10) with Corollary 56 it is evident that the k-support norm is a Θ-norm where $\Theta = \left\{\theta \in \mathbb{R}^d_{++} \;\middle|\; \theta = \sum_{g\in\mathcal{G}_k} \lambda_g 1_g,\ \lambda \in \Delta^{|\mathcal{G}_k|-1}\right\}$, which by symmetry can be expressed as $\Theta = \left\{\theta \;\middle|\; 0 < \theta_i \leq 1, \sum_{i=1}^d \theta_i \leq k\right\}$. Hence, we see that the k-support norm is a special case of the box-norm.

Despite the complicated form of (3.10), Argyriou et al. (2012) observe that the dual norm has a simple formulation, namely the ℓ_2-norm of the k largest components,

$$\|u\|_{(k),*} = \sqrt{\sum_{i=1}^{k} (|u|_i^{\downarrow})^2}, \quad u \in \mathbb{R}^d, \tag{3.11}$$

where $|u|^{\downarrow}$ is the vector obtained from u by reordering its components so that they are non-increasing in absolute value. Note from equation (3.11) that for the cases $k = 1$ and $k = d$, the dual norm is equal to the ℓ_∞-norm and ℓ_2-norm, respectively. It follows that the k-support norm includes the ℓ_1-norm and ℓ_2-norm as special cases.

We now provide a different argument illustrating that the k-support norm belongs to the family of box-norms using the dual norm. We first derive the dual box-norm.

Proposition 58. *The dual box-norm is given by*

$$\|u\|^2_{\mathrm{box},*} = a\|u\|_2^2 + (b-a)\left(\|u\|^2_{(k),*} + (\rho - k)(|u|^{\downarrow}_{k+1})^2\right), \tag{3.12}$$

where $\rho = \frac{c-da}{b-a}$ and k is the largest integer not exceeding ρ.

Proof. We need to solve problem (3.2). We make the change of variable $\phi_i = \frac{\theta_i - a}{b-a}$ and observe that the constraints on θ induce the constraint set $\left\{\phi \in (0,1]^d \mid \sum_{i=1}^d \phi_i \leq \rho\right\}$, where $\rho = \frac{c-da}{b-a}$. Furthermore $\sum_{i=1}^d \theta_i u_i^2 = a\|u\|_2^2 + (b-a)\sum_{i=1}^d \phi_i u_i^2$. The result then follows by taking the supremum over ϕ. □

We see from equation (3.12) that the dual norm decomposes into a weighted combination of the ℓ_2-norm, the k-support norm and a residual term, which vanishes if $\rho = k \in \mathbb{N}_d$. For the rest of this chapter we assume this holds, which loses little generality. This choice is equivalent to requiring that $c = (b-a)k + da$, which is the case considered by Jacob et al. (2009a) in the context of multitask clustering, where $k+1$ is interpreted as the number of clusters and d as the number of tasks. We return to this case in Section 4.2, where we introduce the spectral k-support norm and explain in detail its relationship to the cluster norm.

Observe that if $a = 0, b = 1$, and $\rho = k$, the dual box-norm (3.12) coincides with dual k-support norm in equation (3.11). We conclude that if

$$\Theta = \left\{\theta \in \mathbb{R}^d \;\middle|\; 0 < \theta_i \leq 1,\ \sum_{i=1}^d \theta_i \leq k\right\}$$

then the Θ-norm coincides with the k-support norm.

3.2.1 Properties of the norms

In this section we illustrate a number of properties of the box-norm and the connection to the k-support norm. The first result follows as a special case of Proposition 55.

Corollary 59. *If $0 < a < b$ and $c = (b-a)k + da$, for $k \in \mathbb{N}_d$, then it holds that*

$$\|w\|_{\text{box}} = \inf\left\{\sum_{g\in\mathcal{G}_k}\sqrt{\sum_{i\in g}\frac{v_{g,i}^2}{b} + \sum_{i\notin g}\frac{v_{g,i}^2}{a}} \;\middle|\; v_g \in \mathbb{R}^d,\ \sum_{g\in\mathcal{G}_k} v_g = w\right\},\quad w \in \mathbb{R}^d.$$

Furthermore, the unit ball of the norm is given by the convex hull of the set

$$\bigcup_{g\in\mathcal{G}_k}\left\{w \in \mathbb{R}^d \;\middle|\; \sum_{i\in g}\frac{w_i^2}{b} + \sum_{i\notin g}\frac{w_i^2}{a} \leq 1\right\}. \tag{3.13}$$

Notice in Equation (3.13) that if $b = 1$, then as a tends to zero, we obtain the expression of the k-support norm (3.10), recovering in particular the support constraints. If a is small and positive, the support constraints are not imposed, however most of the weight for each v_g tends to be concentrated on supp(g). Hence, Corollary 59 suggests that if $a \ll b$ then the box-norm regularizer will encourage vectors w whose dominant components are a subset of a union of a small number of groups $g \in \mathcal{G}_k$.

Our next result links two Θ-norms whose parameter sets are related by a linear transformation with positive coefficients.

Lemma 60. *Let Θ be a convex bounded subset of the positive orthant in $\mathbb{R}^d$, and let $\Phi = \{\phi \in \mathbb{R}^d \mid \phi_i = \alpha + \beta\theta_i, \theta \in \Theta\}$, where $\alpha, \beta > 0$. Then*

$$\|w\|_\Phi^2 = \min_{z \in \mathbb{R}^d} \left\{ \frac{1}{\alpha}\|w - z\|_2^2 + \frac{1}{\beta}\|z\|_\Theta^2 \right\}.$$

Proof. We consider the definition of the norm $\|\cdot\|_\Phi$ in (3.1). We have

$$\|w\|_\Phi^2 = \inf_{\phi \in \Phi} \sum_{i=1}^d \frac{w_i^2}{\phi_i} = \inf_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\alpha + \beta\theta_i}, \tag{3.14}$$

where we have made the change of variable $\phi_i = \alpha + \beta\theta_i$. Next we observe that

$$\min_{z \in \mathbb{R}^d} \left\{ \frac{1}{\alpha}\|w - z\|_2^2 + \frac{1}{\beta}\|z\|_\Theta^2 \right\} = \min_{z \in \mathbb{R}^d} \inf_{\theta \in \Theta} \left\{ \sum_{i=1}^d \frac{(w_i - z_i)^2}{\alpha} + \frac{z_i^2}{\beta\theta_i} \right\} = \inf_{\theta \in \Theta} \sum_{i=1}^d \frac{w_i^2}{\alpha + \beta\theta_i}, \tag{3.15}$$

where we interchanged the order of the minimum and the infimum and solved for z componentwise, setting $z_i = \frac{\beta\theta_i w_i}{\alpha + \beta\theta_i}$. The result now follows by combining equations (3.14) and (3.15). □

In Section 3.2 we characterized the k-support norm as a special case of the box-norm. Conversely, Lemma 60 allows us to interpret the box-norm as a perturbation of the k-support norm with a quadratic regularization term.

Proposition 61. *Let $\|\cdot\|_{\text{box}}$ be the box-norm on $\mathbb{R}^d$ with parameters $0 < a < b$ and $c = k(b - a) + da$, for $k \in \mathbb{N}_d$, then*

$$\|w\|_{\text{box}}^2 = \min_{z \in \mathbb{R}^d} \left\{ \frac{1}{a}\|w - z\|_2^2 + \frac{1}{b - a}\|z\|_{(k)}^2 \right\}. \tag{3.16}$$

Proof. The result directly follows from Lemma 60 for $\Theta = \{\theta \in \mathbb{R}^d \mid 0 < \theta_i \leq 1, \sum_{i=1}^d \theta_i \leq k\}$, $\alpha = a$ and $\beta = b - a$. □

Lemma 60 and Proposition 61 can further be interpreted using the Moreau envelope, which we introduced in Definition 2.8. Recall that for $\rho > 0$ and $f : \mathbb{R}^d \to]-\infty, \infty]$ which is proper, lower semi-continuous, the *Moreau envelope* of f with parameter ρ is defined as

$$e_\rho f(w) = \inf_{z \in \mathbb{R}^d} \left\{ f(z) + \frac{1}{2\rho}\|w - z\|_2^2 \right\}.$$

Lemma 60 therefore says that $\beta\|\cdot\|_\Phi^2$ is a Moreau-envelope of $\|\cdot\|_\Theta^2$ with parameter $\rho = \frac{\alpha}{2\beta}$ whenever Φ is defined as $\Phi = \alpha + \beta\Theta$, $\alpha, \beta > 0$. In particular we see from (3.16) that the squared box-norm, scaled by a factor of $(b-a)$, is a Moreau envelope of the squared k-support norm as we have

$$(b-a)\|w\|_{\text{box}}^2 = \min_{z\in\mathbb{R}^d}\left\{\|z\|_{(k)}^2 + \frac{1}{2\rho}\|w-z\|_2^2\right\} =: e_\rho f(w), \tag{3.17}$$

where $f(w) = \|w\|_{(k)}^2$ and $\rho = \frac{1}{2}\frac{a}{b-a}$.

Proposition 61 further allows us to decompose the solution to a vector learning problem using the squared box-norm into two components with particular structure. Specifically, consider the regularization problem

$$\min_{w\in\mathbb{R}^d} \|Xw-y\|_2^2 + \lambda\|w\|_{\text{box}}^2 \tag{3.18}$$

with data $X \in \mathbb{R}^{n\times d}$ and response $y \in \mathbb{R}^n$. Using Proposition 61 and setting $w = u+z$, we see that (3.18) is equivalent to

$$\min_{u,z\in\mathbb{R}^d}\left\{\|X(u+z)-y\|_2^2 + \frac{\lambda}{a}\|u\|_2^2 + \frac{\lambda}{b-a}\|z\|_{(k)}^2\right\}. \tag{3.19}$$

Furthermore, if $(\hat{u}, \hat{z})$ solves problem (3.19) then $\hat{w} = \hat{u}+\hat{z}$ solves problem (3.18). The solution $\hat{w}$ can therefore be interpreted as the superposition of a vector which has small ℓ_2 norm, and a vector which has small k-support norm, with the parameter a regulating these two components. Specifically, as a tends to zero, in order to prevent the objective from blowing up, $\hat{u}$ must also tend to zero and we recover k-support norm regularization. Similarly, as a tends to b, $\hat{z}$ vanishes and we have a simple ridge regression problem.

A further consequence of Proposition 61 is the differentiability of the squared box-norm.

Proposition 62. *If $a > 0$ the squared box-norm is differentiable on $\mathbb{R}^d$ and its gradient*

$$\nabla(\|\cdot\|_{\text{box}}^2) = \frac{2}{a}\left(Id - \text{prox}_{\rho\|\cdot\|_{(k)}^2}\right)$$

is Lipschitz continuous with parameter $\frac{2}{a}$.

Proof. Letting $f(w) = \|w\|_{(k)}^2$, $\rho = \frac{1}{2}\frac{a}{b-a}$, by (3.17) we have $e_\rho f(w) = (b-a)\|w\|_{\text{box}}^2$. The result follows directly from Bauschke and Combettes (2011, Prop. 12.29), as f is convex and continuous on $\mathbb{R}^d$ and the gradient is given as $\nabla(e_\rho f) = \frac{1}{\rho}(\text{Id} - \text{prox}_{\rho f})$. □

Proposition 62 establishes that the square of the box-norm is differentiable and its smoothness is controlled by the parameter a. Furthermore, the gradient can be determined from the proximity operator, which we compute in Section 3.3.

3.2.2 Geometry of the norms

In this section, we briefly investigate the geometry of the box-norm. Figure 3.2 depicts the unit balls for the k-support norm in $\mathbb{R}^3$ for various parameter values, setting $b = 1$ throughout. For $k = 1$ and $k = 3$ we recognize the ℓ_1 and ℓ_2 balls respectively. For $k = 2$ the unit ball retains characteristics of both norms, and in particular we note the discontinuities along each of x, y and z planes, as in the case of the ℓ_1 norm.

Figure 3.3 depicts the unit balls for the box-norm for a range of values of a and k, with $c = (b-a)k + da$. We see that, in general, the balls increase in volume with each of a and k, holding the other parameter fixed. Comparing the k-support norm ($k = 1$), that is the ℓ_1-norm, and the box-norm ($k = 1$, $a = 0.15$), we see that the parameter a smooths out the sharp edges of the ℓ_1 norm. This is also visible when comparing the k-support ($k = 2$) and the box ($k = 2$, $a = 0.15$). This illustrates the smoothing effect of the parameter a, as suggested by Proposition 62.

We can gain further insight into the shape of the unit balls of the box-norm from Corollary 59. Equation (3.13) shows that the primal unit ball is the convex hull of ellipsoids in $\mathbb{R}^d$, where for each group g the semi-principle axis along dimension i has length $\sqrt{b}$ if $i \in g$, and length $\sqrt{a}$ if $i \notin g$. Similarly, the unit ball of the dual box-norm is the intersection of ellipsoids in $\mathbb{R}^d$ where for each group g the semi-principle axis in dimension i has length $1/\sqrt{b}$ if $i \in g$, and length $1/\sqrt{a}$ if $i \notin g$ (see also Equation B.8 in Appendix B). It is instructive to further consider the effect of the parameter a on the unit balls for fixed k. To this end, recall that since $c = (b-a)k + da$, when $k = d$ we have $c = bd$. In this case, for all values of a in $(0, b]$, the objective in (3.1) is attained by setting $\theta_i = b$ for all i, and we recover the ℓ_2-norm, scaled by $1/\sqrt{b}$, for the primal box-norm. Similarly in (3.2), the dual norm gives rise to the ℓ_2-norm, scaled by $\sqrt{b}$. In the remainder of this section we therefore only consider the cases $k \in \{1, 2\}$ in $\mathbb{R}^3$

For $k = 1$, $\mathcal{G}_k = \{\{1\}, \{2\}, \{3\}\}$. The unit ball of the primal box-norm is the convex hull of the ellipsoids defined by

$$\frac{w_1^2}{b} + \frac{w_2^2}{a} + \frac{w_3^2}{a} = 1, \quad \frac{w_1^2}{a} + \frac{w_2^2}{b} + \frac{w_3^2}{a} = 1, \quad \text{and} \quad \frac{w_1^2}{a} + \frac{w_2^2}{a} + \frac{w_3^2}{b} = 1, \tag{3.20}$$

and the unit ball of the dual box-norm is the intersection of the ellipsoids defined by

$$\frac{w_1^2}{b^{-1}} + \frac{w_2^2}{a^{-1}} + \frac{w_3^2}{a^{-1}} = 1, \quad \frac{w_1^2}{a^{-1}} + \frac{w_2^2}{b^{-1}} + \frac{w_3^2}{a^{-1}} = 1, \quad \text{and} \quad \frac{w_1^2}{a^{-1}} + \frac{w_2^2}{a^{-1}} + \frac{w_3^2}{b^{-1}} = 1. \tag{3.21}$$

For $k = 2$, $\mathcal{G}_k = \big\{\{1\}, \{2\}, \{3\}, \{1,2\}, \{2,3\}, \{1,3\}\big\}$. The unit ball of the primal box-norm is the convex hull of the ellipsoids defined by (3.20) in addition to the following

$$\frac{w_1^2}{b}+\frac{w_2^2}{b}+\frac{w_3^2}{a}=1, \quad \frac{w_1^2}{a}+\frac{w_2^2}{b}+\frac{w_3^2}{b}=1, \quad \text{and} \quad \frac{w_1^2}{b}+\frac{w_2^2}{a}+\frac{w_3^2}{b}=1, \tag{3.22}$$

and the unit ball of the dual box-norm is the intersection of the ellipsoids defined by (3.21) in addition to the following

$$\frac{w_1^2}{b^{-1}}+\frac{w_2^2}{b^{-1}}+\frac{w_3^2}{a^{-1}}=1, \quad \frac{w_1^2}{a^{-1}}+\frac{w_2^2}{b^{-1}}+\frac{w_3^2}{b^{-1}}=1, \quad \text{and} \quad \frac{w_1^2}{b^{-1}}+\frac{w_2^2}{a^{-1}}+\frac{w_3^2}{b^{-1}}=1. \tag{3.23}$$

For the primal norm, note that since $b>a$, each of the ellipsoids in (3.20) is entirely contained within one of those defined by (3.22), hence when taking the convex hull we need only consider the latter set. Similarly for the dual norm, since $\frac{1}{b}<\frac{1}{a}$, each of the ellipsoids in (3.21) is contained within one of those defined by (3.23), hence when taking the intersection we need only consider the latter set.

Figures 3.4 and 3.5 depict the constituent ellipses for various parameter values for the primal and dual norms. As a tends to zero the ellipses become degenerate. For $k=1$, taking the convex hull we recover the ℓ_1-norm unit ball in the primal norm, and taking the intersection we recover the ℓ_∞-norm unit ball in the dual norm. As a tends to 1 we recover the ℓ_2 norm in both the primal and the dual.

3.3 Computation of the Norm and the Proximity Operator

In this section, we compute the norm and the proximity operator of the squared box-norm by explicitly solving the optimization problem (3.1). We also specialize our results to the k-support norm and comment on the improvement with respect to the method by Argyriou et al. (2012). Recall that, for every vector $w\in\mathbb{R}^d$, $|w|^\downarrow$ denotes the vector obtained from w by reordering its components so that they are non-increasing in absolute value.

Theorem 63. *For every $w\in\mathbb{R}^d$ it holds that*

$$\|w\|_{\text{box}}^2=\frac{1}{b}\|w_Q\|_2^2+\frac{1}{p}\|w_I\|_1^2+\frac{1}{a}\|w_L\|_2^2, \tag{3.24}$$

where $w_Q=\big(|w|_1^\downarrow,\dots,|w|_q^\downarrow\big)$, $w_I=\big(|w|_{q+1}^\downarrow,\dots,|w|_{d-\ell}^\downarrow\big)$, $w_L=\big(|w|_{d-\ell+1}^\downarrow,\dots,|w|_d^\downarrow\big)$, q and ℓ are the unique integers in $\{0,\dots,d\}$ that satisfy $q+\ell\le d$,

$$\frac{|w_q|}{b}\ge\frac{1}{p}\sum_{i=q+1}^{d-\ell}|w_i|>\frac{|w_{q+1}|}{b}, \qquad \frac{|w_{d-\ell}|}{a}\ge\frac{1}{p}\sum_{i=q+1}^{d-\ell}|w_i|>\frac{|w_{d-\ell+1}|}{a}, \tag{3.25}$$

$p=c-qb-\ell a$ and we have defined $|w_0|=\infty$ and $|w_{d+1}|=0$. Furthermore, the minimizer θ

Figure 3.2: Unit balls of the k-support norm for $k \in \{1,2,3\}$.

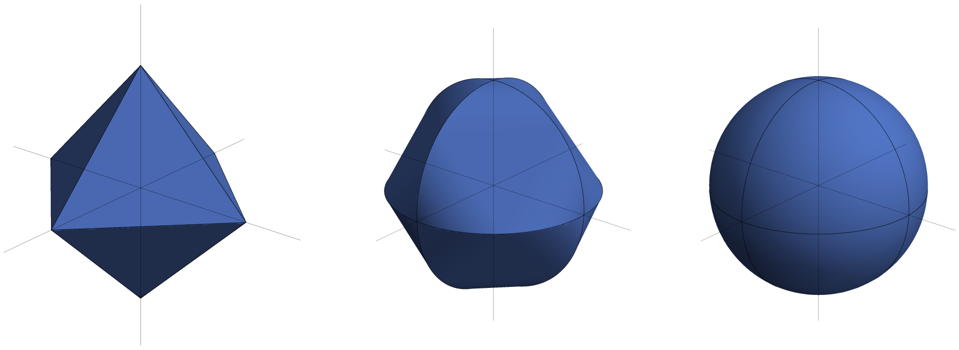


Figure 3.3: Unit balls of the box-norm, $(k,a) \in \{(1,0.15),(2,0.15),(2,0.40)\}$.

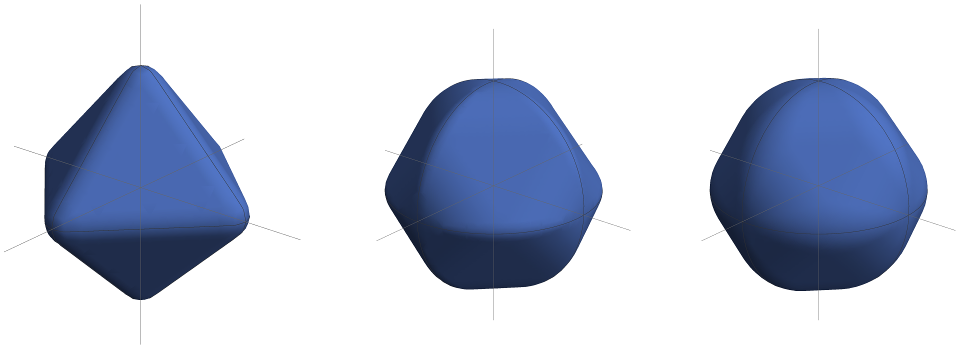


Figure 3.4: Primal box-norm component ellipsoids, $(k,a) \in \{(1,0.15),(2,0.15),(2,0.40)\}$.

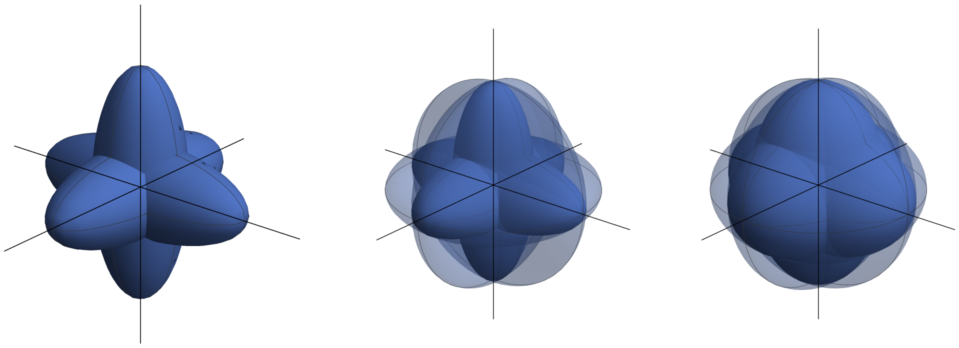
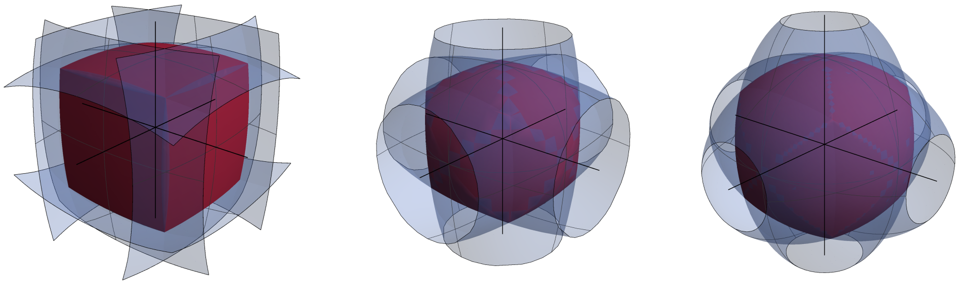


Figure 3.5: Dual box-norm unit balls and ellipsoids, $(k,a) \in \{(1,0.15),(2,0.15),(2,0.40)\}$. For $k=2$, only 3 tightest ellipsoids are shown.



has the form

$$\theta_i = \begin{cases} b, & \text{if } i \in \{1,\dots,q\}, \\ p\frac{|w_i|}{\sum_{j=q+1}^{d-\ell}|w_j|}, & \text{if } i \in \{q+1,\dots,d-\ell\}, \\ a, & \text{otherwise}. \end{cases}$$

Proof. (Sketch) We need to solve the optimization problem

$$\inf_{\theta} \left\{ \sum_{i=1}^{d} \frac{w_i^2}{\theta_i} \;\middle|\; a \le \theta_i \le b, \sum_{i=1}^{d} \theta_i \le c \right\}. \tag{3.26}$$

We assume without loss of generality that the w_i are ordered nonincreasing in absolute values, and it follows that at the optimum the θ_i are also ordered nonincreasing. We further assume that $w_i \neq 0$ for all i and $c \le db$, so the sum constraint will be tight at the optimum, however this can be generalized. The Lagrangian is given by

$$L(\theta,\alpha) = \sum_{i=1}^{d} \frac{w_i^2}{\theta_i} + \frac{1}{\alpha^2}\left(\sum_{i=1}^{d}\theta_i - c\right)$$

where $1/\alpha^2$ is a strictly positive multiplier to be chosen such that $S(\alpha) := \sum_{i=1}^{d}\theta_i(\alpha) = c$. We can then solve the original problem by minimizing the Lagrangian over the constraint $\theta \in [a,b]^d$. Due to the decoupling effect of the multiplier we can solve the simplified problem componentwise, obtaining the solution

$$\theta_i = \theta_i(\alpha) = \min(b, \max(a, \alpha|w_i|)) \tag{3.27}$$

where $S(\alpha) = c$. The minimizer has the form $\theta = (b,\dots,b,\theta_{q+1},\dots,\theta_{d-\ell},a,\dots,a)$, where q,ℓ are determined by the value of α. From $S(\alpha) = c$ we get $\alpha = p/(\sum_{i=q+1}^{d-\ell}|w_i|)$. The value of the norm in (3.24) follows by substituting θ into the objective. Finally, by construction we have $\theta_q \ge b > \theta_{q+1}$ and $\theta_{d-\ell} > a \ge \theta_{d-\ell+1}$, which give rise to the conditions in (3.25). □

Theorem 63 suggests two methods for computing the box-norm. First, we can find α such that $S(\alpha) = c$; this value uniquely determines θ in (3.27), and the norm follows by substitution into the objective in (3.26). Alternatively, we identify q and ℓ that jointly satisfy (3.25) and we compute the norm using (3.24). Taking advantage of the structure of θ in the former method leads to a computation time that is $\mathcal{O}(d \log d)$.

Theorem 64. *The computation of the box-norm can be completed in $\mathcal{O}(d \log d)$ time.*

Proof. Following Theorem 63, we need to determine α^* to satisfy the coupling constraint $S(\alpha^*) = c$. Each component θ_i is a piecewise linear function in the form of a step function with a constant positive slope between the values $a/|w_i|$ and $b/|w_i|$. Let $\{\alpha^i\}_{i=1}^{2d}$ be the set of the $2d$ critical points, where the α^i are ordered nondecreasing. The function $S(\alpha)$ is a nondecreasing piecewise linear function with at most $2d$ critical points. We can find α^* by first sorting the points $\{\alpha^i\}$, finding α^i and α^{i+1} such that

$$S(\alpha^i) \leq c \leq S(\alpha^{i+1})$$

by binary search, and then interpolating α^* between the two points. Sorting takes $\mathcal{O}(d \log d)$. Computing $S(\alpha^i)$ at each step of the binary search is $\mathcal{O}(d)$, so $\mathcal{O}(d \log d)$ overall. Given α^i and α^{i+1}, interpolating α^* is $\mathcal{O}(1)$, so the overall algorithm is $\mathcal{O}(d \log d)$ as claimed. □

The k-support norm is a special case of the box-norm, and as a direct corollary of Theorem 63 and Theorem 64, we recover Argyriou et al. (2012, Proposition 2.1).

Corollary 65. *For $w \in \mathbb{R}^d$, and $k \leq d$,*

$$\|w\|_{(k)} = \left(\sum_{j=1}^{q} (|w|_j^{\downarrow})^2 + \frac{1}{k-q} \Big(\sum_{j=q+1}^{d} |w_j^{\downarrow}| \Big)^2 \right)^{\frac{1}{2}},$$

where q is the unique integer in $\{0, k-1\}$ satisfying

$$|w_q| \geq \frac{1}{k-q} \sum_{j=q+1}^{d} |w_j| > |w_{q+1}|, \tag{3.28}$$

and we have defined $w_0 = \infty$. Furthermore, the norm can be computed in $\mathcal{O}(d \log d)$ time.

3.3.1 Proximity operator

Recall from Section 2.2.5 that proximal gradient methods are efficient optimization methods to solve Tikhonov regularization problems of the form

$$\min_w f(w) + \lambda g(w), \quad w \in \mathbb{R}^d,$$

where f is a convex loss function with Lipschitz continuous gradient, $\lambda > 0$ is a regularization parameter, and g is a convex function. We now use the infimum formulation of the box-norm to derive the proximity operator of the squared norm. We note that recently Chatterjee et al. (2014) showed that the proximity operator of the vector k-support norm can be computed in $O(d \log d)$ time. Here we directly follow Argyriou et al. (2012) and consider the squared k-support norm.

Theorem 66. *The proximity operator of the square of the box-norm at point $w \in \mathbb{R}^d$ with* *parameter $\frac{\lambda}{2}$ is given by* $\mathrm{prox}_{\frac{\lambda}{2}\|\cdot\|^2_{\mathrm{box}}}(w) = \left(\frac{\theta_1 w_1}{\theta_1+\lambda}, \ldots, \frac{\theta_d w_d}{\theta_d+\lambda}\right)$*, where*

$$\theta_i = \begin{cases} b, & \text{if } \alpha|w_i| - \lambda > b, \\ \alpha|w_i| - \lambda, & \text{if } b \geq \alpha|w_i| - \lambda \geq a, \\ a, & \text{if } a > \alpha|w_i| - \lambda, \end{cases}$$

and α is chosen such that $S(\alpha) := \sum_{i=1}^d \theta_i(\alpha) = c$. Furthermore, the computation of the proximity operator can be completed in $\mathcal{O}(d \log d)$ time.

Proof. Using the infimum formulation of the norm, we solve

$$\min_{x\in\mathbb{R}^d} \inf_{\theta\in\Theta} \left\{ \frac{1}{2}\sum_{i=1}^d (x_i - w_i)^2 + \frac{\lambda}{2}\sum_{i=1}^d \frac{x_i^2}{\theta_i} \right\}.$$

We can exchange the order of the optimization and solve for x first. The problem is separable and a direct computation yields that $x_i = \frac{\theta_i w_i}{\theta_i+\lambda}$. Discarding a multiplicative factor of $\lambda/2$, and noting that the infimum is attained, the problem in θ becomes

$$\min_{\theta} \left\{ \sum_{i=1}^d \frac{w_i^2}{\theta_i + \lambda} \;\Big|\; a \leq \theta_i \leq b, \sum_{i=1}^d \theta_i \leq c \right\}.$$

Note that this is the same as computing a box-norm in accordance with Proposition 61. Specifically, this is exactly like problem (3.26) after the change of variable $\theta_i' = \theta_i + \lambda$. The remaining part of the proof then follows in a similar manner to the proof of Theorem 63. □

Algorithm 2 illustrates the computation of the proximity operator for the squared box-norm in $\mathcal{O}(d \log d)$ time. This includes the k-support as a special case, where we let a tend to zero, and set $b = 1$ and $c = k$, which improves upon the complexity of the $\mathcal{O}(d(k + \log d))$ computation provided in Argyriou et al. (2012), and we illustrate the improvement empirically in Table 3.1. We summarize this in the following corollary.

Corollary 67. *The proximity operator of the square of the k-support norm at point w with parameter $\frac{\lambda}{2}$ is given by $\mathrm{prox}_{\frac{\lambda}{2}\|\cdot\|^2_{\Theta}}(w) = x$, where $x_i = \frac{\theta_i w_i}{\theta_i+\lambda}$, and*

$$\theta_i = \begin{cases} 1, & \text{if } \alpha|w_i| > \lambda + 1, \\ \alpha|w_i| - \lambda, & \text{if } \lambda + 1 \geq \alpha|w_i| \geq \lambda \\ 0, & \text{if } \lambda > \alpha|w_i|, \end{cases}$$

where α is chosen such that $S(\alpha) = k$. Furthermore, the proximity operator can be computed in $\mathcal{O}(d \log d)$ time.

Algorithm 2 Computation of $x = \mathrm{prox}_{\frac{\lambda}{2}\|\cdot\|^2_{\mathrm{box}}}(w)$.

```
Require: parameters $a$, $b$, $c$, $\lambda$.
  1. Sort points $\{\alpha^i\}_{i=1}^{2d} = \left\{\frac{a+\lambda}{|w_j|}, \frac{b+\lambda}{|w_j|}\right\}_{j=1}^{d}$ such that $\alpha^i \leq \alpha^{i+1}$;
  2. Identify points $\alpha^i$ and $\alpha^{i+1}$ such that $S(\alpha^i) \leq c$ and $S(\alpha^{i+1}) \geq c$ by binary search;
  3. Find $\alpha^*$ between $\alpha^i$ and $\alpha^{i+1}$ such that $S(\alpha^*) = c$ by linear interpolation;
  4. Compute $\theta_i(\alpha^*)$ for $i = 1 \ldots, d$;
  5. Return $x_i = \frac{\theta_i w_i}{\theta_i + \lambda}$ for $i = 1 \ldots, d$.
```

Table 3.1 empirically verifies the improved performance of the proximity operator computation of Algorithm 2.

Table 3.1: Comparison of proximity operator algorithms for the squared k-support norm (time in seconds), $k = 0.05d$. Algorithm 1 is the method in Argyriou et al. (2012), and Algorithm 2 is our method.

d	1,000	2,000	4,000	8,000	16,000	32,000
Algorithm 1	0.0443	0.1567	0.5907	2.3065	9.0080	35.6199
Algorithm 2	0.0011	0.0016	0.0026	0.0046	0.0101	0.0181

Argyriou et al. (2012) demonstrated the good estimation properties of the vector k-support norm compared to the Lasso and the elastic net. In the following chapter we will extend the norm, along with the box-norm, to the matrix setting. As we will discuss, Algorithm 2 can be applied to matrix learning problems with the induced spectral norms, and we will use this to present a range of numerical experiments with the spectral norms for matrix completion and multitask learning problems.

3.4 The Θ p-Norm

We end this chapter by describing a generalization of the Θ-norm. The primal and dual norms of Equations (3.1) and (3.2) are a special case of the following more general framework. For $p \in [1, \infty]$, define the generalized perspective function (Dacorogna and Maréchal, 2008; Maréchal, 2005) $\phi_p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ by $\phi_p(x) = x^p$ and note that ϕ_p is invertible on $\mathbb{R}_{++}$, and for $p = \infty$ we define its value by continuity.

Proposition 68. *Let $\Theta \subset \mathbb{R}^d_{++}$ be convex and bounded, and let $p \in [1, \infty]$ and let $q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Define the following for $w, u \in \mathbb{R}^d$*

$$\|w\| = \inf_{\theta \in \Theta} \phi_p^{-1}\left(\sum_{i=1}^d \theta_i \phi_p\left(\frac{|w_i|}{\theta_i}\right)\right), \tag{3.29}$$

$$\|u\|_* = \sup_{\theta \in \Theta} \phi_q^{-1}\left(\sum_{i=1}^d \theta_i \phi_q(|u_i|)\right). \tag{3.30}$$

Then the expressions define norms, and furthermore they are dual to each other.

We refer to the expression (3.29) as the Θ p-norm, and (3.30) as the dual Θ p-norm. Clearly we recover the standard Θ-norm when $p = q = 2$.

Proof. The proof of Proposition 68 follows that of Proposition 51, see Appendix B. □

Lemma 54 can be generalized to the Θ p-norm.

Lemma 69 (Scaling). *Let Θ be a convex bounded subset of the open positive orthant, let $p, q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$, let $c > 0$, and define $\tilde{\Theta} = c^q \Theta$. Then the set $\tilde{\Theta}$ defines a primal and dual norm, and there hold $\|\cdot\|_{\tilde{\Theta}} = \frac{1}{c}\|\cdot\|_{\Theta}$ and $\|\cdot\|_{\tilde{\Theta}*} = c\|\cdot\|_{\Theta*}$.*

Proof. Following the reasoning in Lemma 54, the primal and dual norms are well defined. The scaling result follows by noting that with respect to each $\frac{1}{\theta_i}$ (respectively θ_i), the primal (respectively dual) Θ p-norm is a homogeneous function of order $\frac{1}{q}$, using the fact that $\frac{p-1}{p} = \frac{1}{q}$ in the case of the primal norm. □

We provide an example of a Θ p-norm when we introduce the (k, p)-support norm in Chapter 5.

3.5 Discussion

In this chapter we introduced the Θ-norm, we showed that the k-support and box-norms belong to the framework, and we presented a number of results relating to the norms. In the following chapter we turn our attention to matrices, and we show that several properties carry over in a reasonable manner. We then investigate empirically the performance of the penalties in numerical experiments.

Chapter 4

Orthogonally invariant spectral norms

In the previous chapter we introduced the Θ-norm which can be used as a regularizer to impose structure in the solution to a vector learning problem. A special case of this penalty is the k-support norm, which subsumes the ℓ_1-norm used in the lasso method among other sparse learning methods. In this chapter we move to the domain of matrices and we discuss spectral regularizers for low rank and multitask learning problems. By a classical result by von Neumann, vector norms which are symmetric gauge functions induce orthogonally invariance matrix norms. In particular the k-support norm and box-norms of Chapter 3 give rise to spectral norms that exhibit good performance in low rank and multitask learning matrix problems, and we study them in this chapter.

The characteristic properties of matrix norms translate in a natural manner from the respective vector norms. In particular, the unit ball of spectral k-support norm is the convex hull of the set of matrices of rank no greater than k, and Frobenius norm bounded by one. In addition, we observe that the spectral box-norm is essentially equivalent to the cluster norm introduced by Jacob et al. (2009a) for multitask clustering, which in turn can be interpreted as a perturbation of the spectral k-support norm. Moreover, our computation of the vector box-norm and its proximity operator extends naturally to the spectral case, which allows us to use proximal gradient methods to solve regularization problems using the cluster norm. Finally, we provide a method to use the centered versions of the penalties, which are important in applications (see e.g. Evgeniou et al., 2007; Jacob et al., 2009a), and we provide numerical experiments using the norms.

This chapter is organized as follows. In Section 4.1 we extend the k-support and box norms to orthogonally invariant matrix norms. In Section 4.2, we review multitask learning. In Section 4.3 we review clustered multitask learning and we discuss the cluster norm introduced by Jacob et al. (2009a). In Section 4.4 we provide a method for solving the resulting matrix regularization problem using “centered" norms. Finally, in Section 4.5, we apply the norms to matrix learning problems on a number of simulated and real datasets.

4.1 Orthogonally Invariant Θ-Norms

Recall from Section 2.5.4 that an orthogonally invariant matrix norm $\|\cdot\|$ is invariant to left and right multiplication by orthogonal matrices, that is for $X \in \mathbb{R}^{d\times m}$,

$$\|W\| = \|AWB^\top\| = \|\Sigma_W\|,$$

where $A \in \mathbf{O}^d$, $B \in \mathbf{O}^m$ are orthogonal, and Σ_W is the matrix of singular values of W. A classical result by von Neumann (Theorem 5 in Chapter 2) states that a matrix norm is orthogonally invariant if and only if it is of the form $\|W\| = g(\sigma(W))$, where $\sigma(W)$ is the vector formed by the singular values of W, and g is a symmetric gauge function, that is, a vector norm which is invariant under permutations and sign changes (Von Neumann, 1937).

Examples of norms which are related by this property include the ℓ_1-norm and trace norm, the ℓ_2-norm and the Frobenius norm, and the ℓ_∞-norm and the spectral norm (the largest singular value of a matrix).

The relationship extends beyond the computation of the norm. In particular, we highlight that the proximity operator of an orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ can be computed using von Neumann's trace inequality (Theorem 7 in Chapter 2). Specifically, it is given by

$$\mathrm{prox}_{\|\cdot\|}(W) = U\mathrm{diag}\big(\mathrm{prox}_g(\sigma(W))\big)V^\top, \quad W \in \mathbb{R}^{d\times m}, \tag{4.1}$$

where U and V are the matrices formed by the left and right singular vectors of W, see e.g. Argyriou et al. (2011, Prop. 3.1) and Lewis (2003). It follows that whenever we can efficiently compute the proximity operator of a symmetric gauge function, we can compute the proximity operator of the induced spectral norm at the additional cost of computing the singular value decomposition, and we will make use of this fact below.

In Chapter 3 we introduced the Θ-norm and discussed several examples. Using von Neumann's theorem we can extend certain instances to matrices.

Lemma 70. *If Θ is a convex bounded subset of the strictly positive orthant in $\mathbb{R}^d$ that is invariant under permutations, then $\|\cdot\|_\Theta$ is a symmetric gauge function.*

Proof. Let $g(w) = \|w\|_\Theta$. We need to show that g is a norm that is invariant under permutations and sign changes. By Proposition 51, g is a norm, so it remains to show that $g(w) = g(Pw)$ for every permutation matrix P, and $g(Jw) = g(w)$ for every diagonal matrix J with entries ± 1. The former property holds since the set Θ is permutation invariant. The latter property holds because the objective function in (3.1) involves the squares of the components of w. □

4.1.1 The spectral k-support norm and spectral box-norm

Lemma 70 establishes that both the k-support norm and the box-norm are spectral gauge functions. We can therefore extend both to orthogonally invariant norms that we term the spectral k-support norm and the spectral box-norm respectively, and which we write (with some abuse of notation) as $\|W\|_{(k)} = \|\sigma(W)\|_{(k)}$ and $\|W\|_{\mathrm{box}} = \|\sigma(W)\|_{\mathrm{box}}$. We note that since the k-support norm subsumes the ℓ_1 and ℓ_2-norms for $k = 1$ and $k = d$ respectively, the corresponding spectral k-support norms are equal to the trace and Frobenius norms respectively.

A number of properties of the vector norms translate in the natural manner to the matrix norms. We first characterize the unit ball of the spectral k-support norm.

Proposition 71. *The unit ball of the spectral k-support norm is the convex hull of the set of matrices of rank at most k and Frobenius norm no greater than one.*

Proof. (Sketch) We consider the sets

$$T_k = \left\{W \in \mathbb{R}^{d\times m} \;\middle|\; \operatorname{rank}(W) \le k,\ \|W\|_F \le 1\right\}, \quad A_k = \operatorname{co}(T_k),$$

and we apply Lemma 36 to the set A_k to derive a norm, and finally we argue that the norm is a spectral function with the desired properties, see Appendix C. □

Referring to the unit ball characterization of the k-support norm, we note that the restriction on the cardinality of the vectors that define the extreme points of the unit ball naturally extends to a restriction on the rank operator in the matrix setting. Furthermore, as noted by Argyriou et al. (2012), regularization using the k-support norm encourages vectors to be sparse, but less so than the ℓ_1-norm. In matrix regularization problems, Proposition 71 suggests that the spectral k-support norm for $k > 1$ encourages matrices to have low rank, but less so than the trace norm. This is intuitive as the extreme points of the unit ball have rank at most k.

As in the case of the vector norm (Proposition 61), the spectral box-norm (or cluster norm – see below) can be written as a perturbation of the spectral k-support norm with a quadratic term.

Proposition 72. *Let $\|\cdot\|_{\mathrm{box}}$ be a matrix box-norm with parameters a, b, c and let $k = \frac{c-da}{b-a}$. Then*

$$\|W\|_{\mathrm{box}}^2 = \min_{Z\in\mathbb{R}^{d\times m}} \left\{\frac{1}{a}\|W-Z\|_F^2 + \frac{1}{b-a}\|Z\|_{(k)}^2\right\}.$$

Proof. By von Neumann's trace inequality (Theorem 7) we have

$$\begin{aligned}\frac{1}{a}\|W-Z\|_F^2+\frac{1}{b-a}\|Z\|_{(k)}^2 &= \frac{1}{a}\left(\|W\|_F^2+\|Z\|_F^2-2\langle W \mid Z\rangle\right)+\frac{1}{b-a}\|Z\|_{(k)}^2\\ &\geq \frac{1}{a}\left(\|\sigma(W)\|_2^2+\|\sigma(Z)\|_2^2-2\langle\sigma(W)\mid\sigma(Z)\rangle\right)+\frac{1}{b-a}\|\sigma(Z)\|_{(k)}^2\\ &= \frac{1}{a}\|\sigma(W)-\sigma(Z)\|_2^2+\frac{1}{b-a}\|\sigma(Z)\|_{(k)}^2.\end{aligned}$$

Furthermore the inequality is tight if W and Z have the same ordered set of singular vectors. Hence

$$\min_{Z\in\mathbb{R}^{d\times m}}\left\{\frac{1}{a}\|W-Z\|_F^2+\frac{1}{b-a}\|Z\|_{(k)}^2\right\}=\min_{z\in\mathbb{R}^d}\left\{\frac{1}{a}\|\sigma(W)-z\|_2^2+\frac{1}{b-a}\|z\|_{(k)}^2\right\}=\|\sigma(W)\|_{\mathrm{box}}^2,$$

where the last equality follows by Proposition 61. □

In other words, this result shows that the (scaled) squared spectral box-norm can be seen as the Moreau envelope of a squared spectral k-support norm.

Finally, the computational considerations outlined in Section 3.3 can be naturally extended to the matrix setting by using 4.1. Using this result we can employ proximal gradient methods to solve matrix regularization problems using the square of the spectral k-support and box-norms, see Section 4.5.

4.2 Multitask Learning

We now briefly review multitask learning, which we introduced in Chapter 2, and which we return to in the numerical experiments below. This is a framework in which a number of learning tasks are solved concurrently in order to leverage commonalities between them (Baxter, 2000; Romera-Paredes, 2014), with applications in particular in the fields of machine vision (Torralba et al., 2004), natural language processing (Do and Ng, 2005) and biometrics (Sun, 2008). Supervised learning applications frequently involve several learning problems which are not completely independent. For example given a collection of images from a given domain, we could expect a classifier for one object to be related to a classifier for a similar object. A reasonable assumption is that both classifiers take advantage of the underlying features in the input, hence aggregating the training data and solving the problems together should give improved performance, in particular when limited training data is available for each task. In this context, multitask learning is related to transfer learning, or learning to learn (Pan and Yang, 2010), as well as multi-class logistic regression (Wright et al., 2009).

We focus on multitask learning approaches that use spectral regularizers to encode the relationship between the tasks (Evgeniou et al., 2005; Argyriou et al., 2008; Jacob et al., 2009a). Each problem, or task, is represented by a single weight vector, and these are stacked as the columns of a matrix W which is to be learned. By convention we use T to denote the number of tasks, each of which is a vector in $\mathbb{R}^d$. Given training points $(x_1^t, y_1^t), \ldots, (x_n^t, y_n^t) \in \mathbb{R}^d \times \mathbb{R}$ for task t (for simplicity we assume that each task has the same number n of training points), the problem is given by

$$\min_{W \in \mathbb{R}^{d \times T}} \mathcal{L}(W) + \lambda \Omega(W), \tag{4.2}$$

where $W = [w_1, \ldots, w_T]$ is the $d \times T$ matrix whose columns represent the task vectors, $\mathcal{L}$ is the empirical error $\mathcal{L}(W) = \frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} \ell(y_i^t, \langle w_t \mid x_i^t \rangle)$ for some convex loss function ℓ, and Ω is a regularizer which promotes structure between the tasks.

A number of different regularizers have been proposed to promote different relationships between the tasks and we highlight two approaches below.

4.2.1 Hierarchical methods

Given tasks $w_1, \ldots, w_T$, hierarchical methods assume a relationship between the tasks of the form

$$w_t = w_0 + v_t, \tag{4.3}$$

for each t. By encouraging v_t to be small, the task vectors w_t are all closely related to the root component w_0. One approach to this is given by penalizing the variance between the tasks (Evgeniou and Pontil, 2004; Evgeniou et al., 2005; Torralba et al., 2004), using a regularizer of the form

$$\begin{aligned}\Omega(W) &= \min_{w_0} \frac{1}{\lambda T} \sum_{t=1}^{T} \|w_t - w_0\|_2^2 + \frac{1}{1-\lambda} \|w_0\|_2^2 \\ &= \frac{1}{T} \left(\sum_{t=1}^{T} \|w_t\|_2^2 + \frac{1-\lambda}{\lambda} \sum_{t=1}^{T} \left\| w_t - \frac{1}{T} \sum_{s=1}^{T} w_s \right\|_2^2 \right),\end{aligned}$$

where $\lambda \in [0,1]$ is a parameter to be tuned. For small values of λ, the tasks are closely aligned, and as λ increases to 1, the tasks become less closely related.

4.2.2 Sparsity inducing penalties

A priori information may suggest that the weight vectors should be sparse in a given application, and the support sets may be the same or they may vary between the tasks. For example, when identifying how DNA microarrays cause related cancers, a common assumption is that these are due to a small number of sets of genes present in the multiarray (Khan et al., 2001).

A number of matrix $\ell_{p,q}$-norms have been proposed as regularizers, that is

$$\Omega(W) = \|W\|_{p,q} = \left(\sum_{t=1}^{T} \left(\sum_{i=1}^{d} |w_{i,t}|^q \right)^{\frac{p}{q}} \right)^{\frac{1}{p}},$$

for given values of p and q, or alternatively

$$\Omega(W) = \|W^\top\|_{p,q} = \left(\sum_{i=1}^{d} \left(\sum_{t=1}^{T} |w_{i,t}|^q \right)^{\frac{p}{q}} \right)^{\frac{1}{p}}.$$

These norms aggregate the ℓ_q-norms of the columns, respectively rows, of W via the ℓ_p-norms of these values.

In particular setting $q = 1$ promotes sparsity of the task vectors, for different values of p depending on the underlying assumptions. Applying the $\ell_{1,\infty}$-norm to the transpose of the matrix, Turlach et al. (2005) shrink the ℓ_1-norm of the maximum value of each explanatory variable across all the tasks, that is

$$\|W^\top\|_{1,\infty} = \sum_{i=1}^{d} \max_{t=1}^{T} |w_{i,t}|.$$

Finally, we note that matrix completion can be formulated as an instance of multitask learning (Srebro et al., 2005). In this framework, each user is a task, the input data are indicators for the corresponding items, and the labels are the values of the entries.

4.2.3 Clustering the tasks

A natural assumption that arises in MTL applications is that the tasks are clustered. The cluster norm was introduced by Jacob et al. (2009a) as a means to favour this structure and we review the norm here. Recall that our general approach to multitask learning is based on the regularization problem (4.2)

$$\min_{W \in \mathbb{R}^{d \times T}} \mathcal{L}(W) + \lambda \Omega(W).$$

Jacob et al. (2009a) consider a composite penalty which encourages the tasks to be clustered into $Q < T$ groups. To introduce their setting we require some more notation. Let $\mathcal{J}_q \subset \mathbb{N}_T$ be the set of tasks in cluster $q \in \mathbb{N}_Q$ and let $T_q = |\mathcal{J}_q| \geq 0$ be the number of tasks in cluster q, so that $\sum_{q=1}^{Q} T_q = T$. The clustering uniquely defines the $T \times T$ normalized connectivity matrix M where $M_{st} = \frac{1}{T_q}$ if $s,t \in \mathcal{J}_q$ and $M_{st} = 0$ otherwise. We let $\bar{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$ be the mean weight vector, $\bar{w}_q = \frac{1}{T_q} \sum_{t \in \mathcal{J}_q} w_t$ be the mean weight vector of tasks in cluster q and define the $T \times T$ orthogonal projection matrices $U = 11^\top / T$ and $\Pi = I - U$. Note that

$W\Pi = [w_1 - \bar{w}, \ldots, w_T - \bar{w}]$. Finally, let $r = \min(d, T)$.

Using this notation, we introduce the three seminorms

$$\begin{aligned}
\Omega_{\mathrm{m}}(W) &= T\|\bar{w}\|^2 = \operatorname{tr}\left(WUW^\top\right) \\
\Omega_{\mathrm{b}}(W) &= \sum_{q=1}^{Q} T_q \|\bar{w}_q - \bar{w}\|^2 = \operatorname{tr}\left(W(M-U)W^\top\right) \\
\Omega_{\mathrm{w}}(W) &= \sum_{q=1}^{Q} \sum_{t \in \mathcal{J}_q} \|w_t - \bar{w}_q\|^2 = \operatorname{tr}\left(W(I-M)W^\top\right),
\end{aligned}$$

each of which captures a different aspect of the clustering: Ω_{m} penalizes the total *mean* of the weight vectors, Ω_{b} measures how close to each other the clusters are (*between* cluster variance), and Ω_{w} measures the compactness of the clusters (*within* cluster variance). Scaling the three penalties by positive parameters ϵ_{m}, ϵ_{b}, and ϵ_{w} respectively, we obtain the composite penalty $\epsilon_{\mathrm{m}}\Omega_{\mathrm{m}} + \epsilon_{\mathrm{b}}\Omega_{\mathrm{b}} + \epsilon_{\mathrm{w}}\Omega_{\mathrm{w}}$. The first term Ω_{m} does not depend on the connectivity matrix M, and it can be included in the error term. The remaining two terms depend on M, which in general may not be known *a-priori*. Jacob et al. (2009a) propose to learn the clustering by minimizing with respect to matrix M, under the assumption that $\epsilon_{\mathrm{w}} \geq \epsilon_{\mathrm{b}}$. This assumption is reasonable as we care more about enforcing a small variance of parameters within the clusters than between them. Using the elementary properties that $M - U = M\Pi = \Pi M \Pi$ and $I - M = (I - M)\Pi = \Pi(I - M)\Pi$ and letting $\widetilde{M} = M\Pi$, we rewrite

$$\epsilon_{\mathrm{b}}\Omega_{\mathrm{b}}(W) + \epsilon_{\mathrm{w}}\Omega_{\mathrm{w}}(W) = \operatorname{tr}\left(W\Pi\big(\epsilon_{\mathrm{b}}\widetilde{M} + \epsilon_{\mathrm{w}}(I - \widetilde{M})\big)\Pi W^\top\right) = \operatorname{tr}\left(W\Pi\Sigma^{-1}\Pi W^\top\right) \tag{4.4}$$

where we have defined $\Sigma^{-1} = \epsilon_{\mathrm{b}}\widetilde{M} + \epsilon_{\mathrm{w}}(I - \widetilde{M})$. Since $\widetilde{M}$ is an orthogonal projection, the matrix Σ is well defined and we have

$$\Sigma = (\epsilon_{\mathrm{b}}^{-1} - \epsilon_{\mathrm{w}}^{-1})\widetilde{M} + \epsilon_{\mathrm{w}}^{-1} I. \tag{4.5}$$

The expression in the right hand side of equation (4.4) is jointly convex in W and Σ (see e.g. Boyd and Vandenberghe, 2004), however the set of matrices Σ defined by equation (4.5), generated by letting $\widetilde{M} = M\Pi$ vary, is nonconvex, because M takes values on a nonconvex set. To address this, Jacob et al. (2009a) relax the constraint on matrix $\widetilde{M}$ to the set $\{0 \preceq \widetilde{M} \preceq I,\ \operatorname{tr}\widetilde{M} \leq Q - 1\}$. This in turn induces the convex constraint set for Σ

$$\mathcal{S}_{Q,T} = \left\{\Sigma \in \mathbb{R}^{T\times T} \,\middle|\, \Sigma = \Sigma^\top,\ \epsilon_{\mathrm{w}}^{-1} I \preceq \Sigma \preceq \epsilon_{\mathrm{b}}^{-1} I,\ \operatorname{tr}\Sigma \leq (\epsilon_{\mathrm{b}}^{-1} - \epsilon_{\mathrm{w}}^{-1})(Q-1) + \epsilon_{\mathrm{w}}^{-1} T\right\}. \tag{4.6}$$

In summary Jacob et al. (2009a) arrive at the optimization problem

$$\min_{W \in \mathbb{R}^{d \times T}} \bar{\mathcal{L}}(W) + \lambda \|W\Pi\|_{\mathrm{cl}}^2 \tag{4.7}$$

where $\bar{\mathcal{L}}(W) = \mathcal{L}(W) + \lambda\epsilon_{\mathrm{m}} \operatorname{tr}(WUW^\top)$ and $\|\cdot\|_{\mathrm{cl}}$ is the *cluster norm* defined by the equation

$$\|W\|_{\mathrm{cl}} = \sqrt{\inf_{\Sigma \in \mathcal{S}_{Q,T}} \operatorname{tr}(W\Sigma^{-1}W^\top)}. \tag{4.8}$$

4.3 The Cluster Norm and the Spectral Box-Norm

We now discuss the cluster norm in the context of the spectral box-norm. Jacob et al. (2009a) state that the cluster norm of W equals what we have termed the spectral box-norm, with parameters $a = \epsilon_{\mathrm{w}}{}^{-1}$, $b = \epsilon_{\mathrm{b}}^{-1}$ and $c = (T-Q+1)\epsilon_{\mathrm{w}}^{-1} + (Q-1)\epsilon_{\mathrm{b}}{}^{-1}$, that is referring to (4.6) we get the constraint set

$$\mathcal{S}_{a,b,c} = \left\{ \Sigma \in \mathbb{R}^{T \times T} \;\middle|\; \Sigma = \Sigma^\top,\ aI \preceq \Sigma \preceq bI,\ \operatorname{tr}\Sigma \le c \right\}, \tag{4.9}$$

and we recover Equation (2.28) from Section 2.5.9 (Jalali et al., 2016). Here we prove this fact. Denote by $\lambda_i(\cdot)$ the eigenvalues of a matrix which we write in non increasing order $\lambda_1(\cdot) \ge \lambda_2(\cdot) \ge \ldots \ge \lambda_d(\cdot)$. Note that if θ_i are the eigenvalues of Σ then $\theta_i = \lambda_{d-i+1}(\Sigma^{-1})$. We have that

$$\operatorname{tr}(\Sigma^{-1}W^\top W) \ge \sum_{i=1}^{r} \lambda_{d-i+1}(\Sigma^{-1})\lambda_i(W^\top W) = \sum_{i=1}^{r} \frac{\sigma_i^2(W)}{\theta_i}$$

where the inequality follows by Lemma 8 (see Chapter 2) for $A = \Sigma^{-1}$ and $B = W^\top W \succeq 0$. Since this inequality is attained whenever $\Sigma = U\operatorname{diag}(\theta)U$, where U are the eigenvectors of $W^\top W$, we see that the cluster norm coincides with the spectral box-norm, that is $\|W\|_{\mathrm{cl}} = \|\sigma(W)\|_\Theta$ for $\Theta = \left\{\theta \in [a,b]^r \;\middle|\; \sum_{i=1}^r \theta_i \le c\right\}$.

In light of our observations in Section 4.1, we also see that the spectral k-support norm is a special case of the cluster norm, where we let a tend to zero, $b = 1$ and $c = k$, where $k = Q - 1$. More importantly the cluster norm is a perturbation of the spectral k-support norm. Moreover, as we outlined in Section 4.1, the methods to compute the norm and its proximity operator (cf. Theorems 63 and 66) can directly be applied to the cluster norm using von Neumann's trace inequality.

4.4 Optimization with Centered Spectral Θ-Norms

Centering a matrix has been shown to improve learning in other multitask learning problems, for example Evgeniou et al. (2005) reported improved results using the trace norm. It is therefore valuable to address the problem of how to solve a regularization problem of the type

$$\min_{W\in\mathbb{R}^{d\times T}} \bar{\mathcal{L}}(W)+\lambda\|W\Pi\|_{\Theta}^{2} \tag{4.10}$$

in which the regularizer is applied to the matrix $W\Pi=[w_1-\bar{w},\dots,w_T-\bar{w}]$. To this end, let Θ be a bounded and convex subset of $\mathbb{R}^r_{++}$ which is invariant under permutation. We have already noted that the function defined, for every $W\in\mathbb{R}^{d\times T}$, as

$$\|W\|_{\Theta} := \|\sigma(W)\|_{\Theta},$$

is an orthogonally invariant norm. In particular, problem (4.10) includes regularization with the centered cluster norm outlined above.

Note that right multiplication by the centering operator Π, is invariant to a translation of the columns of the matrix by a fixed vector, that is, for every $z\in\mathbb{R}^d$, we have $[w_1+z,\dots,w_T+z]\Pi=W\Pi$. The quadratic term $\epsilon_{\mathrm{m}}\mathrm{tr}(WUW^{\top})$, which is included in the error, implements square norm regularization of the mean of the tasks, which can help to prevent overfitting. However, in the remainder of this section this term plays no role in the analysis, which equally applies to the case that $\epsilon_{\mathrm{m}}=0$.

In order to solve the problem (4.10) with a centered regularizer the following lemma is key.

Lemma 73. *Let $r=\min(d,T)$ and let Θ be a bounded and convex subset of $\mathbb{R}^r_{++}$ which is invariant under permutation. For every $W=[w_1,\dots,w_T]\in\mathbb{R}^{d\times T}$, it holds that*

$$\|W\Pi\|_{\Theta}=\min_{z\in\mathbb{R}^d}\|[w_1-z,\dots,w_T-z]\|_{\Theta}.$$

Proof. (Sketch) We define the set $\Theta^{(T)}=\left\{\Sigma\in\mathbf{S}^T_{++}\;\middle|\;\lambda(\Sigma)\in\Theta\right\}$ and $\Theta^{(d)}=\left\{D\in\mathbf{S}^d_{++}\;\middle|\;\lambda(D)\in\Theta\right\}$ and apply Lemma 8 to get

$$\|W\|_{\Theta}^2\equiv\|\sigma(W)\|_{\Theta}^2=\inf_{\Sigma\in\Theta^{(T)}}\mathrm{tr}\left(\Sigma^{-1}W^{\top}W\right)=\inf_{D\in\Theta^{(d)}}\mathrm{tr}\left(D^{-1}WW^{\top}\right).$$

The result then follows by using the definition of $W\Pi=[w_1-\bar{w},\dots,w_T-\bar{w}]$ and solving the optimization problem, see Appendix C. □

Using this lemma, we rewrite problem (4.7) as

$$\min_{W\in\mathbb{R}^{d\times T}}\min_{z\in\mathbb{R}^d}\bar{\mathcal{L}}(W)+\lambda\|[w_1-z,\dots,w_T-z]\|_{\Theta}.$$

Letting $\boldsymbol{w}_t = w_t - z$, and $\boldsymbol{\mathcal{W}} = [v_1, \dots, v_T]$, we obtain the equivalent problem

$$\min_{(\boldsymbol{\mathcal{W}},z)\in\mathbb{R}^{d\times T}\times\mathbb{R}^d} \bar{\mathcal{L}}(V + z1^\top) + \lambda\|\boldsymbol{\mathcal{W}}\|_\Theta^2. \tag{4.11}$$

This problem is of the form $f(V,z) + \lambda g(V,z)$, where $g(V,z) = \|V\|_\Theta$. Using this formulation, we can directly apply the proximal gradient method using the proximity operator computation for the vector norm, since $\operatorname{prox}_g(V_0, z_0) = (\operatorname{prox}_{\lambda\|\cdot\|_\Theta}(V_0), z_0)$. This observation establishes that, whenever the proximity operator of the spectral Θ-norm is available, we can use proximal gradient methods with minimal additional effort to perform optimization with the corresponding centered spectral Θ-norm. For example, this is the case with the trace norm, the spectral k-support norm and the spectral box-norm or cluster norm.

4.5 Numerical Experiments

Argyriou et al. (2012) demonstrated the good estimation properties of the vector k-support norm compared to the lasso and the elastic net. In this section, we investigate the matrix norms and report on their statistical performance in matrix completion and multitask learning experiments on simulated as well as benchmark real datasets. We also offer an interpretation of the role of the parameters in the box-norm.

We compare the spectral k-support norm (*k-sup*) and the spectral box-norm (*box*) to the baseline trace norm (*trace*) (see e.g. Argyriou et al., 2007a; Mazumder et al., 2010; Srebro et al., 2005; Toh and Yun, 2011), matrix elastic net (*el.net*) (Li et al., 2012) and, in the case of multitask learning, the Frobenius norm (*fr*), which we recall is equivalent to the spectral k-support norm when $k = d$. As we highlighted in Section 4.4, centering a matrix can lead to improvements in learning. For datasets which we expect to exhibit clustering we therefore also apply centered versions of the norms, *c-fr, c-trace, c-el.net, c-k-sup, c-box.*[1]

We report test error and standard deviation, matrix rank (r), and optimal parameter values for k and a, which are determined by validation. We used a t-test to determine the statistical significance of the difference in performance between the regularizers, at a level of $p < 0.001$.

To solve the optimization problem we used an accelerated proximal gradient method (FISTA), (see e.g. Beck and Teboulle, 2009; Nesterov, 2007), using the percentage change in the objective as convergence criterion, with a tolerance of 10^{-5} (10^{-3} for real matrix completion experiments).

[1] As we described in Section 4.3, the cluster norm regularization problem from Jacob et al. (2009a) is equivalent to regularization using the box-norm with a squared ℓ_2 norm of the mean column vector included in the loss function. The centering operator is invariant to constant shifts of the columns, which allows the matrix to have unbounded Frobenius norm when using a centered regularizer. The additional quadratic term regulates this effect and can prevent against overfitting. We tested the effect of the quadratic term on the centered norms, however the impact on performance was only incremental, and it introduced a further parameter requiring validation. On the real datasets in particular, the impact was not significant compared to simple centering, so we do not report on the method below.

As is typical with spectral regularizers such as the trace norm, we found that the spectrum of the learned matrix exhibited a rapid decay to zero. In order to explicitly impose a low rank on the final matrix, we included a thresholding step at the end of the optimization. For the matrix completion experiments, the thresholding level was chosen by validation. Matlab code used in the experiments is available at `http://wwwo.cs.ucl.ac.uk/staff/M.Pontil/software.html`.

4.5.1 Simulated data

Matrix Completion. We applied the norms to matrix completion on noisy observations of low rank matrices. Each $d \times d$ matrix was generated as $W = AB^\top + E$, where $A, B \in \mathbb{R}^{d \times r}$, $r \ll d$, and the entries of A, B and E were set to be i.i.d. standard Gaussian. We set $d = 100$, $r \in \{5, 10\}$ and we sampled uniformly a percentage $\rho \in \{10\%, 10\%, 20\%, 30\%\}$ of the entries for training, and used a fixed 10% for validation. Following Mazumder et al. (2010) the error was measured as

$$\text{error} = \frac{\|w_{\text{true}} - w_{\text{predicted}}\|^2}{\|w_{\text{true}}\|^2},$$

and averaged over 100 trials. The results are summarized in Table 4.1. With thresholding, all methods recovered the rank of the true noiseless matrix. The spectral box-norm generated the lowest test errors in all regimes, with the spectral k-support norm a close second, and both were significantly better than trace and elastic net.

Table 4.1: Matrix completion on simulated datasets, without (left) and with (right) thresholding.

dataset	norm	test error	r	k	a	test error	r	k	a
rank 5	trace	0.8184 (0.03)	20	-	-	0.7799 (0.04)	5	-	-
ρ=10%	el.net	0.8164 (0.03)	20	-	-	0.7794 (0.04)	5	-	-
	k-sup	0.8036 (0.03)	16	3.6	-	0.7728 (0.04)	5	4.23	-
	box	0.7805 (0.03)	87	2.9	1.7e-2	0.7649 (0.04)	5	3.63	8.1e-3
rank 5	trace	0.5764 (0.04)	22	-	-	0.5209 (0.04)	5	-	-
ρ=15%	el.net	0.5744 (0.04)	21	-	-	0.5203 (0.04)	5	-	-
	k-sup	0.5659 (0.03)	18	3.3	-	0.5099 (0.04)	5	3.25	-
	box	0.5525 (0.04)	100	1.3	9e3	0.5089 (0.04)	5	3.36	2.7e-3
rank 5	trace	0.4085 (0.03)	23	-	-	0.3449 (0.02)	5	-	-
ρ=20%	el.net	0.4081 (0.03)	23	-	-	0.3445 (0.02)	5	-	-
	k-sup	0.4031 (0.03)	21	3.1	-	0.3381 (0.02)	5	2.97	-
	box	0.3898 (0.03)	100	1.3	9e-3	0.3380 (0.02)	5	3.28	1.9e-3
rank 10	trace	0.6356 (0.03)	27	-	-	0.6084 (0.03)	10	-	-
ρ=20%	el.net	0.6359 (0.03)	27	-	-	0.6074 (0.03)	10	-	-
	k-sup	0.6284 (0.03)	24	4.4	-	0.6000 (0.03)	10	5.02	-
	box	0.6243 (0.03)	89	1.8	9e-3	0.6000 (0.03)	10	5.22	1.9e-3
rank 10	trace	0.3642 (0.02)	36	-	-	0.3086 (0.02)	10	-	-
ρ=30%	el.net	0.3638 (0.02)	36	-	-	0.3082 (0.02)	10	-	-
	k-sup	0.3579 (0.02)	33	5.0	-	0.3025 (0.02)	10	5.13	-
	box	0.3486 (0.02)	100	2.5	9e-3	0.3025 (0.02)	10	5.16	3e-4

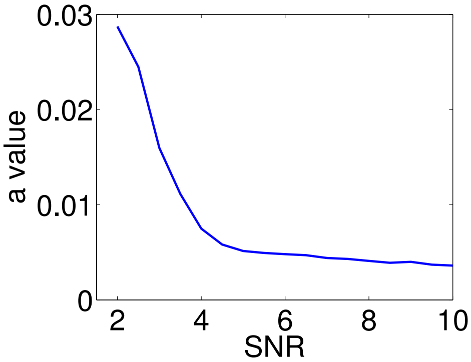


Figure 4.1: Impact of SNR on value of a.

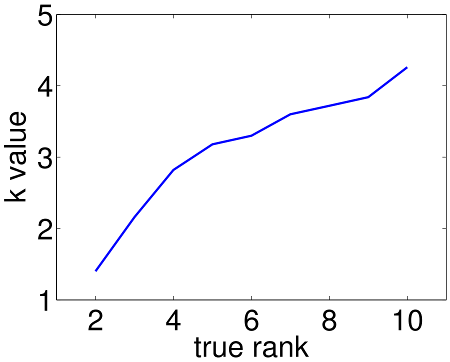
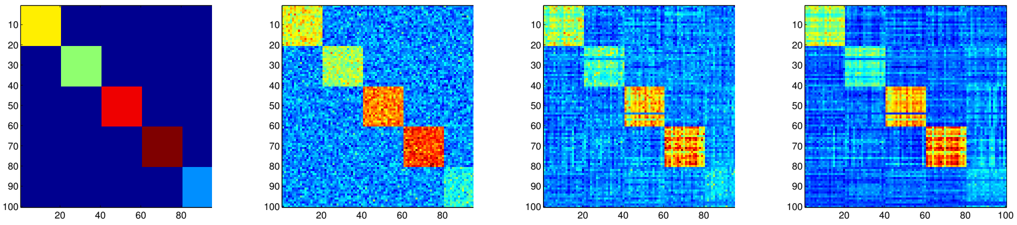


Figure 4.2: Impact of matrix rank on value of k.

Figure 4.3: Clustered matrix and recovered solutions. From left to right: true, noisy, trace norm, box-norm



Role of Parameters. In the same setting we investigated the role of the parameters in the box-norm. As previously discussed, parameter b can be set to 1 without loss of generality. Figure 4.1 shows the optimal value of parameter a chosen by validation for varying signal to noise ratios (SNR), keeping k fixed. We see that for greater noise levels (smaller SNR), the optimal value for a increases, which further suggests that the noise is filtered out by higher values of the parameter. Figure 4.2 shows the optimal value of k chosen by validation for matrices with increasing rank, keeping a fixed, and using the relation $k = \frac{c-da}{b-a}$. We note that as the rank of the matrix increases, the optimal k value increases, which is expected since it is an upper bound on the sum of the singular values.

Clustered Learning. We tested the centered norms on a synthetic dataset which exhibited a clustered structure. We generated a 100×100, rank 5, block diagonal matrix, where the entries of each 20×20 block were set to a random integer chosen uniformly in $\{1, \ldots, 10\}$, with additive noise. Table 4.2 illustrates the results averaged over 100 runs. Within each group of norms, the box-norm and the k-support norm outperformed the trace norm and elastic net, and centering improved performance for all norms. Figure 4.3 illustrates a sample matrix along with the solution found using the box and trace norms.

4.5.2 Real data

Matrix Completion (MovieLens and Jester). In this section we report on the performance

Table 4.2: Clustered block diagonal matrix, before (left) and after (right) thresholding.

dataset	norm	test error	r	k	a	test error	r	k	a
ρ=10%	trace	0.6529 (0.10)	20	-	-	0.6065 (0.10)	5	-	-
	el.net	0.6482 (0.10)	20	-	-	0.6037 (0.10)	5	-	-
	k-sup	0.6354 (0.10)	19	2.72	-	0.5950 (0.10)	5	2.77	-
	box	0.6182 (0.09)	100	2.23	1.9e-2	0.5881 (0.10)	5	2.73	4.3e-3
	c-trace	0.5959 (0.07)	15	-	-	0.5692 (0.07)	5	-	-
	c-el.net	0.5910 (0.07)	14	-	-	0.5670 (0.07)	5	-	-
	c-k-sup	0.5837 (0.07)	14	2.03	-	0.5610 (0.07)	5	1.98	-
	c-box	0.5789 (0.07)	100	1.84	1.9e-3	0.5581 (0.07)	5	1.93	9.7e-3
ρ=15%	trace	0.3482 (0.08)	21	-	-	0.3048 (0.07)	5	-	-
	el.net	0.3473 (0.08)	21	-	-	0.3046 (0.07)	5	-	-
	k-sup	0.3438 (0.07)	21	2.24	-	0.3007 (0.07)	5	2.89	-
	box	0.3431 (0.07)	100	2.05	8.7e-3	0.3005 (0.07)	5	2.57	1.3e-3
	c-trace	0.3225 (0.07)	19	-	-	0.2932 (0.06)	5	-	-
	c-el.net	0.3215 (0.07)	18	-	-	0.2931 (0.06)	5	-	-
	c-k-sup	0.3179 (0.07)	18	1.89	-	0.2883 (0.06)	5	2.36	-
	c-box	0.3174 (0.07)	100	1.90	2.2e-3	0.2876 (0.06)	5	1.92	3.8e-3

of the norms on real datasets. We observe a subset of the (user, rating) entries of a matrix and the task is to predict the unobserved ratings, with the assumption that the true matrix is low rank (or approximately low rank). In the first instance we considered the MovieLens datasets[2]. These consist of user ratings of movies, the ratings are integers from 1 to 5, and all users have rated a minimum number of 20 films. Specifically we considered the following datasets:

- *MovieLens 100k*: 943 users and 1,682 movies, with a total of 100,000 ratings;
- *MovieLens 1M*: 6,040 users and 3,900 movies, with a total of 1,000,209 ratings.

We also considered the Jester [3] datasets, which consist of user ratings of jokes, where the ratings are real values from -10 to 10:

- *Jester 1*: 24,983 users and 100 jokes, all users have rated a minimum of 36 jokes;
- *Jester 2*: 23,500 users and 100 jokes, all users have rated a minimum of 36 jokes;
- *Jester 3*: 24,938 users and 100 jokes, all users have rated between 15 and 35 jokes.

Following Toh and Yun (2011), for MovieLens we uniformly sampled $\rho = 50\%$ of the available entries for each user for training, and for Jester 1, Jester 2 and Jester 3 we sampled 20, 20 and 8 ratings per user respectively, and we again used 10% for validation. The error was measured as normalized mean absolute error,

$$\text{NMAE} = \frac{\|w_{\text{true}} - w_{\text{predicted}}\|^2}{\#\text{observations}/(r_{\max} - r_{\min})},$$

[2] MovieLens datasets are available at *http://grouplens.org/datasets/movielens/*.
[3] Jester datasets are available at *http://goldberg.berkeley.edu/jester-data/*.

Table 4.3: Matrix completion on real datasets, without (left) and with (right) thresholding.

dataset	norm	test error	r	k	a	test error	r	k	a
MovieLens	trace	0.2034	87	-	-	0.2017	13	-	-
100k	el.net	0.2034	87	-	-	0.2017	13	-	-
$\rho = 50\%$	k-sup	0.2031	102	1.00	-	0.1990	9	1.87	-
	box	0.2035	943	1.00	1e-5	0.1989	10	2.00	1e-5
MovieLens	trace	0.1821	325	-	-	0.1790	17	-	-
1M	el.net	0.1821	319	-	-	0.1789	17	-	-
$\rho = 50\%$	k-sup	0.1820	317	1.00	-	0.1782	17	1.80	-
	box	0.1817	3576	1.09	3e-5	0.1777	19	2.00	1e-6
Jester 1	trace	0.1787	98	-	-	0.1752	11	-	-
20 per line	el.net	0.1787	98	-	-	0.1752	11	-	-
	k-sup	0.1764	84	5.00	-	0.1739	11	6.38	-
	box	0.1766	100	4.00	1e-6	0.1726	11	6.40	2e-5
Jester2	trace	0.1767	98	-	-	0.1758	11	-	-
20 per	el.net	0.1767	98	-	-	0.1758	11	-	-
line	k-sup	0.1762	94	4.00	-	0.1746	11	4.00	-
	box	0.1762	100	4.00	2e-6	0.1745	11	4.50	5e-5
Jester 3	trace	0.1988	49	-	-	0.1959	3	-	-
8 per line	el.net	0.1988	49	-	-	0.1959	3	-	-
	k-sup	0.1970	46	3.70	-	0.1942	3	2.13	-
	box	0.1973	100	5.91	1e-3	0.1940	3	4.00	8e-4

where $r_{\max}$ and $r_{\min}$ are upper and lower bounds for the ratings (Toh and Yun, 2011), averaged over 50 runs. The results are outlined in Table 4.3. In the thresholding case, the spectral box-norm and the spectral k-support norm showed the best performance, and in the absence of thresholding, the spectral k-support norm showed slightly improved performance. Comparing to the synthetic datasets, this suggests that the parameter a did not provide any benefit in the absence of noise. We also note that without thresholding our results for trace norm regularization on MovieLens 100k agreed with those in Jaggi and Sulovsky (2010).

Multitask Learning (Lenk and Animals with Attributes). In our final set of experiments we considered two multitask learning datasets, where we expected the data to exhibit clustering. The *Lenk personal computer* dataset (Lenk et al., 1996) consists of 180 ratings of 20 profiles of computers characterized by 14 features (including a bias term). The clustering is suggested by the assumption that users are motivated by similar groups of features. We used the root mean square error of true vs. predicted ratings, normalised over the tasks, averaged over 100 runs. We also report on the Frobenius norm, which in the multitask learning framework corresponds to independent task learning. The results are outlined in Table 4.4. The centered versions of the spectral k-support norm and spectral box-norm outperformed the other penalties in all regimes. Furthermore, the results clearly indicate the importance of centering, as discussed for the trace norm in Evgeniou et al. (2007).

The *Animals with Attributes* dataset (Lampert et al., 2009) consists of 30,475 images of animals from 50 classes. Along with the images, the dataset includes pre-extracted features

Table 4.4: Multitask learning clustering on Lenk dataset.

norm	test error	k	a
fr	3.7931 (0.07)	-	-
trace	1.9056 (0.04)	-	-
el.net	1.9007 (0.04)	-	-
k-sup	1.8955 (0.04)	1.02	-
box	1.8923 (0.04)	1.01	5.5e-3
c-fr	1.8634 (0.08)	-	-
c-trace	1.7902 (0.03)	-	-
c-el.net	1.7897 (0.03)	-	-
c-k-sup	1.7777 (0.03)	1.89	-
c-box	1.7759 (0.03)	1.12	8.6e-3

for each image. The dataset has been analyzed in the context of multitask learning. We followed the experimental protocol from Kang et al. (2011), however we used an updated feature set, and we considered all 50 classes. Specifically, we used the DeCAF feature set provided by Lampert et al. (2009) rather than the SIFT bag of word descriptors. These updated features were obtained through a deep convolutional network and represent each image by a 4,096-dimensional vector (Donahue et al., 2014). As the smallest class size is 92 we selected the first $n = 92$ examples of each of the $T = 50$ classes, used PCA (with centering) on the resulting data matrix to reduce dimensionality ($d = 1,718$) retaining a variance of 95%, and obtained a dataset of size $4,600 \times 1,718$. For each class the examples were split into training, validation and testing datasets, with a split of 50%, 25%, 25% respectively, and we averaged the performance over 50 runs.

We used the logistic loss, yielding the error term

$$\mathcal{L}(W) = \sum_{t=1}^{T} \sum_{i=1}^{Tn} \log\left(1 + \exp(-y_{t,i} \langle w_t \mid x_i \rangle)\right)$$

where $W = [w_1, \ldots, w_T]$, $x_1, \ldots, x_{Tn}$ are the inputs and $y_{t,i} = 1$ if x_i is in class t, and $y_{t,i} = -1$ otherwise.

The predicted class for testing example x was $\operatorname{argmax}_{t=1}^{T} \langle w_t \mid x \rangle$ and the accuracy was measured as the percentage of correctly classified examples, also known as multi-class error. The results without centering are outlined in Table 4.5. The corresponding results with centering showed the same relative performance, but worse overall accuracy, which is reasonable as the data is not expected to be clustered, and we omit the results here.

The spectral k-support and box-norms gave the best results, outperforming the Frobenius norm and the matrix elastic net, which in turn outperformed the trace norm. We highlight that in contrast to the Lenk experiments, the Frobenius norm, corresponding to independent task learning, was competitive. Furthermore, the optimal values of k for the spectral k-support norm and spectral box-norm were high (38 and 33, respectively) relative to the maximum

Table 4.5: Multitask learning clustering on Animals with Attributes dataset, no centering.

norm	test error	k	a
fr	38.3428 (0.74)	-	-
tr	37.4285 (0.76)	-	-
el.net	38.2857 (0.73)	-	-
k-sup	38.8571 (0.71)	37.8	-
box	38.9100 (0.65)	32.8	2.1e-2

rank of 50, corresponding to a relatively high rank solution. The spectral k-support norm and spectral box-norm nonetheless outperformed the other regularizers. Notice also that the spectral k-support norm requires the same number of parameters to be tuned as the matrix elastic net, which suggests that it somehow captures the underlying structure of the data in a more appropriate manner.

We finally note as an aside that using the SIFT bag of words descriptors provided by Lampert et al. (2009), which represent the images as a $2,000$-dimensional histogram of local features, we replicated the results for independent task learning (Frobenius norm regularization) and single-group learning (trace norm regularization) of Kang et al. (2011) for the subset of 20 classes considered in their paper.

4.6 Discussion

In this chapter we extended the vector k-support and box-norms to orthogonally invariant matrix norms, we presented a number of theoretical results as well as experiments showing the performance of the norms as regularizers. In this section we outline a number of extensions to the norms discussed in this chapter. The first two topics we address in subsequent chapters, so we do not elaborate here.

4.6.1 The k-support p-norm

A natural extension of the k-support norm follows by applying a p-norm, rather than the Euclidean norm, in the infimal convolution definition of the k-support norm. We investigate this in Chapter 5.

4.6.2 The tensor k-support norm

The k-support norm can further be extended to low rank tensor learning by restricting the rank of the matricizations of a tensor. We discuss this in Chapter 6.

4.6.3 Kernels

The ideas discussed in this chapter can be used in the context of multiple kernel learning in a natural way (see e.g. Micchelli and Pontil, 2007, and references therein). Let K_j, $j \in \mathbb{N}_s$, be prescribed reproducing kernels on a set X , and H_j the corresponding reproducing kernel

Hilbert spaces with norms $\|\cdot\|_j$. We consider the problem

$$\min\left\{\sum_{i=1}^{n}\ell\Big(y_i,\sum_{\ell=1}^{s}f_\ell(x_i)\Big)+\lambda\big\|\big(\|f_1\|_1,\ldots,\|f_s\|_s\big)\big\|_\Theta^2 \;\Big|\; f_1\in H_1,\ldots,f_s\in H_s\right\}.$$

The choice $\Theta=\left\{\theta\in\mathbb{R}^d \;\middle|\; 0<\theta_i\leq 1,\ \sum_{i=1}^{d}\theta_i\leq k\right\}$, when $k\leq s$, is particularly interesting. It gives rise to a version of multiple kernel learning in which at least k kernels are employed.

4.6.4 Rademacher complexity

We briefly comment on the Rademacher complexity of the spectral k-support norm, namely

$$\mathbb{E}\sup_{\|W\|_{(k)}\leq 1}\frac{1}{Tn}\sum_{t=1}^{T}\sum_{i=1}^{n}\epsilon_i^t\langle w_t \mid x_i^t\rangle$$

where the expectations is taken with respect to i.i.d. Rademacher random variables ϵ_i^t, $i\in\mathbb{N}_n,\ t\in\mathbb{N}_T$ and the x_i^t are either prescribed or random datapoints associated with the different regression tasks. The Rademacher complexity can be used to derive uniform bounds on the estimation error and excess risk bounds (see Bartlett and Mendelson, 2002; Koltchinskii and Panchenko, 2002, for a discussion). Although a complete analysis is beyond the scope of this thesis, we remark that the Rademacher complexity of the unit ball of the spectral k-support is a factor of $\sqrt{k}$ larger than the Rademacher complexity bound for the trace norm provided in Maurer and Pontil (Proposition 6 2013). This follows from the fact that the dual spectral k-support norm is bounded by $\sqrt{k}$ times the operator norm. Of course the unit ball of the spectral k-support norm contains the unit ball of the trace norm, so the associated excess risk bounds need to be compared with care.

Chapter 5

The (k,p)-Support Norm

In Chapter 3 we discussed the Θ-norm, defined as the square root of an infimum of quadratics, and which captures a wide range of penalties by tuning the parameter set Θ. In Section 3.4 we generalized this notion to the Θ p-norm, which features a p-th power in the objective of the primal norm, and the corresponding Hölder conjugate power in the dual norm. A particular instance of the Θ-norm that we studied was the k-support norm. In the matrix setting in particular, the norm has good estimation properties, as we showed for low rank matrix completion and multitask learning problems in Section 4.5.

In this chapter, we introduce the *(k,p)-support norm*, which we show is an instance of the Θ p-norm. An interesting property of the norm is that it interpolates between the ℓ_1 norm (for $k=1$) and the ℓ_p-norm (for $k=d$). It follows that varying both k and p the norm allows one to learn sparse vectors which exhibit different patterns of decay in the non-zero elements. The (k,p)-support norm is a symmetric gauge function, hence it induces the orthogonally invariant *spectral (k,p)-support norm*. This interpolates between the trace norm (for $k=1$) and the Schatten p-norms (for $k=d$), and its unit ball has a simple geometric interpretation as the convex hull of matrices of rank no greater than k and Schatten p-norm no greater than one. This suggests that the new norm favors low rank structure and the effect of varying p allows different patterns of decay patterns in the spectrum, allowing the model to adapt to a variety of decay patterns of the singular values. We explicitly compute the norm, we provide a conditional gradient method to solve regularization problems with the penalty, and we derive an efficient algorithm to compute the Euclidean projection on the unit ball in the case $p=\infty$. In numerical experiments, we show that allowing p to vary significantly improves performance over the spectral k-support norm on various matrix completion benchmarks, and better captures the spectral decay of the underlying model.

This chapter is organized as follows. In Section 5.1 we introduce the (k,p)-support norm. In Section 5.2 we compute the norm and in Section 5.3 we describe how to solve Ivanov regularization problems using the vector and matrix (k,p)-support norms. Finally in Section

5.4 we apply the spectral norm to matrix completion problems.

5.1 Extending the k-Support Norm

Recall that for every $k \in \mathbb{N}_d$, the k-support norm $\|\cdot\|_{(k)}$ is defined as the norm whose unit ball is given by

$$\mathrm{co}\big\{w \in \mathbb{R}^d \;\big|\; \mathrm{card}(w) \le k, \|w\|_2 \le 1\big\}, \tag{5.1}$$

that is, the convex hull of the set of vectors of cardinality at most k and ℓ_2-norm no greater than one (Argyriou et al., 2012). Equivalently, the k-support norm of a vector $w \in \mathbb{R}^d$ can be expressed as an infimal convolution (Rockafellar, 1970, p. 34),

$$\|w\|_{(k)} = \inf_{(v_g)} \left\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_2 \;\Big|\; \sum_{g \in \mathcal{G}_k} v_g = w \right\}, \tag{5.2}$$

where $\mathcal{G}_k$ is the collection of all subsets of $\mathbb{N}_d$ containing at most k elements and the infimum is over all vectors $v_g \in \mathbb{R}^d$ such that $\mathrm{supp}(v_g) \subset g$, for $g \in \mathcal{G}_k$. Furthermore, the dual norm has a simple form, namely the ℓ_2-norm of the k largest components,

$$\|u\|_{(k),*} = \sqrt{\sum_{i=1}^{k} (|u|_i^{\downarrow})^2}, \quad u \in \mathbb{R}^d, \tag{5.3}$$

where $|u|^{\downarrow}$ is the vector obtained from u by reordering its components so that they are nonincreasing in absolute value (Argyriou et al., 2012).

Recall that a norm is orthogonally invariant if and only if it is of the form $\|W\| = g(\sigma(W))$, where $\sigma(W)$ is the vector formed by the singular values of W, and g is a symmetric gauge function (Von Neumann, 1937). As the set $\mathcal{G}_k$ includes all subsets of size k, $\|\cdot\|_{(k)}$ is a symmetric gauge function and we used this fact to introduce the spectral k-support norm for matrices in Chapter 4, by defining $\|W\|_{(k)} = \|\sigma(W)\|_{(k)}$, for $W \in \mathbb{R}^{d\times m}$.

We can now introduce the (k,p)-support norm. This follows by applying the ℓ_p-norm, rather than the Euclidean norm, in the infimal convolution definition of the norm (5.2).

Definition 74. *Let $k \in \mathbb{N}_d$ and $p \in [1,\infty]$. The (k,p)-support norm of a vector $w \in \mathbb{R}^d$ is defined as*

$$\|w\|_{(k,p)} = \inf_{(v_g)} \left\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_p \;\Big|\; \sum_{g \in \mathcal{G}_k} v_g = w \right\}. \tag{5.4}$$

where the infimum is over all vectors $v_g \in \mathbb{R}^d$ such that $\mathrm{supp}(v_g) \subset g$, *for $g \in \mathcal{G}_k$.*

Let us note that the norm is well defined. Indeed, positivity, homogeneity and non degeneracy are immediate. To prove the triangle inequality, let $w, w' \in \mathbb{R}^d$. For any $\epsilon > 0$ there exist $\{v_g\}$ and $\{v'_g\}$ such that $w = \sum_g v_g$, $w' = \sum_g v'_g$, $\sum_g \|v_g\|_p \le \|w\|_{(k,p)} + \epsilon/2$, and $\sum_g \|v'_g\|_p \le \|w'\|_{(k,p)} + \epsilon/2$. As $\sum_g v_g + \sum_g v'_g = w + w'$, we have

$$\begin{aligned} \|w + w'\|_{(k,p)} &\le \sum_g \|v_g\|_p + \sum_g \|v'_g\|_p \\ &\le \|w\|_{(k,p)} + \|w'\|_{(k,p)} + \epsilon, \end{aligned}$$

and the result follows by letting ϵ tend to zero.

Note that, since a convex set is equivalent to the convex hull of its extreme points, Definition 74 implies that the unit ball of the (k,p)-support norm, denoted by C_k^p, is given by the convex hull of the set of vectors with cardinality no greater than k and ℓ_p-norm no greater than 1, that is

$$C_k^p = \mathrm{co}\big\{w \in \mathbb{R}^d \;\big|\; \mathrm{card}(w) \le k, \|w\|_p \le 1\big\}. \tag{5.5}$$

Definition 74 gives the norm as the solution of a variational problem. Its explicit computation is not straightforward in the general case, however for $p = 1$ the unit ball (5.5) does not depend on k and is always equal to the ℓ_1 unit ball. Thus, the $(k,1)$-support norm is always equal to the ℓ_1-norm, and we do not consider further this case in this section. Similarly, for $k = 1$ we recover the ℓ_1-norm for all values of p. For $p = \infty$, from the definition of the dual norm it is not difficult to show that $\|\cdot\|_{(k,p)} = \max\{\|\cdot\|_\infty, \|\cdot\|_1/k\}$. We return to this in Section 5.2 when we describe how to compute the norm for all values of p.

Note further that in Equation (5.4), as p tends to ∞, the ℓ_p-norm of each v_g is increasingly dominated by the largest component of v_g. As the variational formulation tries to identify vectors v_g with small aggregate ℓ_p-norm, this suggests that higher values of p encourage each v_g to tend to a vector whose k entries are equal. In this manner varying p allows us adjust the degree to which the components of vector w can be clustered into (possibly overlapping) groups of size k.

As in the case of the k-support norm, the dual (k,p)-support norm has a simple expression. Recall that the dual norm of a vector $u \in \mathbb{R}^d$ is defined by the optimization problem

$$\|u\|_{(k,p),*} = \max\big\{\langle u \mid w\rangle \;\big|\; \|w\|_{(k,p)} = 1\big\}. \tag{5.6}$$

Proposition 75. *If $p \in (1,\infty]$ then the dual (k,p)-support norm is given by*

$$\|u\|_{(k,p),*} = \left(\sum_{i \in I_k} |u_i|^q\right)^{\frac{1}{q}}, \quad u \in \mathbb{R}^d,$$

where $q = p/(p-1)$ and $I_k \subset \mathbb{N}_d$ is the set of indices of the k largest components of u in absolute value. Furthermore, if $p \in (1,\infty)$ and $u \in \mathbb{R}^d \backslash \{0\}$ then the maximum in (5.6) *is attained for*

$$w_i = \begin{cases} \operatorname{sign}(u_i)\left(\frac{|u_i|}{\|u\|_{(k,p),*}}\right)^{\frac{1}{p-1}} & \textit{if } i \in I_k, \\ 0 & \textit{otherwise}. \end{cases} \tag{5.7}$$

If $p = \infty$ the maximum is attained for

$$w_i = \begin{cases} \operatorname{sign}(u_i) & \textit{if } i \in I_k, u_i \neq 0, \\ \lambda_i \in [-1,1] & \textit{if } i \in I_k, u_i = 0, \\ 0 & \textit{otherwise}. \end{cases}$$

Proof. The proof follows by computing the dual norm, and using Hölder's inequality. See Appendix D. □

Note that for $p = 2$ we recover the dual of the k-support norm in (5.3).

5.1.1 The (k,p)-support norm as a Θ-norm

In Section 3.4 we presented the Θ p-norm. Using that framework we can express the (k,p)-support norm as follows.

Lemma 76 (The (k,p)-support norm is a Θ p-norm.). *Let $p,q \in [1,\infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Let $\Theta = \left\{\theta \in \mathbb{R}^d \;\middle|\; 0 < \theta_i \leq 1, \sum_{i=1}^d \theta_i \leq k\right\}$. The (k,p)-support norm and its dual norm are respectively given, for $w,u \in \mathbb{R}^d$, by*

$$\|w\|_{(k,p)} = \left(\inf_{\theta \in \Theta} \sum_{i=1}^d \frac{|w_i|^p}{\theta_i^{p-1}}\right)^{\frac{1}{p}}, \tag{5.8}$$

$$\|u\|_{(k,p),*} = \left(\sup_{\theta \in \Theta} \sum_{i=1}^d \theta_i |u_i|^q\right)^{\frac{1}{q}}. \tag{5.9}$$

Proof. We first show that the dual norms are equivalent. Assume without loss of generality that the components of u are nonnegative, and ordered nonincreasing, that is $u_1 \geq \ldots \geq u_d \geq 0$. By inspection the solution to (5.9) is given by choosing $\theta_1 = \ldots = \theta_k = 1$, and setting the remaining components to zero, and we recover the ℓ_q-norm of the k largest components of u, that is the dual of the (k,p)-support norm. It follows that that primal norms are also equivalent and we are done. □

This characterization shows that the (k,p)-support norm fits naturally within the framework of the Θ p-norms.

5.1.2 The spectral (k,p)-support norm

From Definition 74 it is clear that the (k,p)-support norm is a symmetric gauge function. This follows since $\mathcal{G}_k$ contains all groups of cardinality k and the ℓ_p-norms only involve absolute values of the components. Hence we can define the spectral (k,p)-support norm as

$$\|W\|_{(k,p)} = \|\sigma(W)\|_{(k,p)}, \quad W \in \mathbb{R}^{d\times m}.$$

Since the dual of any orthogonally invariant norm is given by $\|\cdot\|_* = \|\sigma(\cdot)\|_*$, see e.g. Lewis (1995), we conclude that the dual spectral (k,p)-support norm is given by

$$\|Z\|_{(k,p),*} = \|\sigma(Z)\|_{(k,p),*}, \quad Z \in \mathbb{R}^{d\times m}.$$

The next result characterizes the unit ball of the spectral (k,p)-support norm. Due to the relationship between an orthogonally invariant norm and its corresponding symmetric gauge function, we see that the cardinality constraint for vectors generalizes in a natural manner to the rank operator for matrices.

Proposition 77. *The unit ball of the spectral (k,p)-support norm is the convex hull of the set of matrices of rank at most k and Schatten p-norm no greater than one.*

In particular, if $p=\infty$, the dual vector norm is given by $u\in\mathbb{R}^d$, by $\|u\|_{(k,\infty),*} = \sum_{i=1}^k |u|_i^\downarrow$. Hence, for any $Z\in\mathbb{R}^{d\times m}$, the dual spectral norm is given by $\|Z\|_{(k,\infty),*} = \sum_{i=1}^k \sigma_i(Z)$, that is the sum of the k largest singular values, which is also known as the Ky-Fan k-norm, see e.g. Bhatia (1997).

Proof. The proof follows by the Minkowski functional construction of Lemma 36, see Appendix D. □

5.2 Computing the Norm

In this section we compute the norm, illustrating how it interpolates between the ℓ_1 and ℓ_p-norms.

Theorem 78. *Let $p\in(1,\infty)$. For every $w\in\mathbb{R}^d$, and $k\le d$, it holds that*

$$\|w\|_{(k,p)} = \left[\sum_{i=1}^{\ell} (|w|_i^\downarrow)^p + \left(\frac{\sum_{i=\ell+1}^d |w|_i^\downarrow}{\sqrt[q]{k-\ell}}\right)^p\right]^{\frac{1}{p}} \tag{5.10}$$

where $\frac{1}{p}+\frac{1}{q}=1$, and for $k=d$, we set $\ell=d$, otherwise ℓ is the largest integer in $\{0,\ldots,k-1\}$

satisfying

$$(k-\ell)|w|^{\downarrow}_{\ell} \geq \sum_{i=\ell+1}^{d} |w|^{\downarrow}_{i}. \tag{5.11}$$

Furthermore, the norm can be computed in $\mathcal{O}(d\log d)$ time.

Proof. (Sketch) We compute $\|w\|_{(k,p)}$ as $\max\left\{\sum_{i=1}^{d} u_i w_i \;\middle|\; \|u\|_{(k,p),*} \leq 1\right\}$, or equivalently

$$\max\left\{\sum_{i=1}^{k-1} u_i z_i + u_k \sum_{i=k}^{d} z_i \;\middle|\; \sum_{i=1}^{k} u_i^q \leq 1, u_1 \geq \cdots \geq u_k\right\}, \tag{5.12}$$

where $z_i = |w|^{\downarrow}_{i}$, and we obtain a k-dimensional problem. We use Hölder's inequality to find that the components of the maximizer u satisfy $u_i = (z_i/M_{k-1})^{p-1}$ if $i \leq k-1$, and $u_k = \left(\sum_{i=\ell+1}^{d} z_i/M_{k-1}\right)^{p-1}$, where for every $\ell \in \{0,\ldots,k-1\}$, M_ℓ denotes the r.h.s. in equation (5.10). The ordering constraints are satisfied if inequality (5.11) holds for $\ell = k-1$. If this inequality is true we are done, otherwise we set $u_k = u_{k-1}$ and solve the lower dimensional equivalent of problem (5.12) in the same fashion. Solving the general case in this manner leads to (5.11). The computational complexity follows from the fact that M_ℓ is a nondecreasing function of ℓ and using binary search. □

Note that for $k = d$ we recover the ℓ_p-norm and for $p = 2$ we recover the result in Section 3.3.

Remark 79 (Computation of the norm for $p \in \{1,\infty\}$)**.** *Since the norm $\|\cdot\|_{(k,p)}$ computed above for $p \in (1,\infty)$ is continuous in p, the special cases $p = 1$ and $p = \infty$ can be derived by a limiting argument. We readily see that for $p = 1$ the norm does not depend on k and it is always equal to the ℓ_1-norm, in agreement with our observation in the previous section. For $p = \infty$ we obtain that $\|w\|_{(k,\infty)} = \max(\|w\|_\infty, \|w\|_1/k)$.*

5.3 Optimization

In this section, we describe how to solve Ivanov regularization problems using the vector and matrix (k,p)-support norms. We consider the constrained optimization problem

$$\min\left\{f(w) \;\middle|\; \|w\|_{(k,p)} \leq \alpha\right\}, \tag{5.13}$$

where w is in $\mathbb{R}^d$ or $\mathbb{R}^{d\times m}$, $\alpha > 0$ is a regularization parameter and the error function f is assumed to be convex and continuously differentiable.

A convenient tool to solve problem (5.13) is provided by the *Frank-Wolfe* method (Frank and Wolfe, 1956), see also Jaggi (2013) for a recent account. The method is outlined in

Algorithm 3 Frank-Wolfe algorithm.

```
Choose $w^{(0)}$ such that $\|w^{(0)}\|_{(k,p)} \le \alpha$
for $t = 0,\dots,T$ do
    Compute $g := \nabla f(w^{(t)})$
    Compute $s := \operatorname{argmin}\big\{\langle s \mid g\rangle \;\big|\; \|s\|_{(k,p)} \le \alpha\big\}$
    Update $w^{(t+1)} := (1-\gamma)w^{(t)} + \gamma s$, for $\gamma := \frac{2}{t+2}$
end for
```

Algorithm 3, and it has worst case convergence rate $\mathcal{O}(1/T)$. The key step of the algorithm is to solve the subproblem

$$\operatorname{argmin}\big\{\langle s \mid g\rangle \;\big|\; \|s\|_{(k,p)} \le \alpha\big\}, \tag{5.14}$$

where $g = \nabla f(w^{(t)})$, that is the gradient of the objective function at the t-th iteration. This problem involves computing a subgradient of the dual norm at g. It can be solved exactly and efficiently as a consequence of Proposition 75. We discuss here the vector case and postpone the discussion of the matrix case to Section 5.3.2. By symmetry of the ℓ_p-norm, problem (5.14) can be solved in the same manner as the maximum in Proposition 75 and the solution is given by $s_i = -\alpha w_i$, where w_i is given by (5.7) .

Specifically, letting $I_k \subset \mathbb{N}_d$ be the set of indices of the k largest components of g in absolute value, for $p \in (1,\infty)$ we have

$$s_i = \begin{cases} \alpha \operatorname{sign}(g_i) \left(\frac{|g_i|}{\|g\|_{(k,p),*}}\right)^{\frac{1}{p-1}}, & \text{if } i \in I_k \\ 0, & \text{if } i \notin I_k \end{cases} \tag{5.15}$$

and, for $p = \infty$ we choose the subgradient

$$s_i = \begin{cases} \alpha \operatorname{sign}(g_i) & \text{if } i \in I_k,\; g_i \ne 0, \\ 0 & \text{otherwise.} \end{cases} \tag{5.16}$$

5.3.1 Projection operator

An alternative method to solve (5.13) in the vector case is to consider the equivalent Tikhonov problem

$$\min\Big\{f(w) + \delta_{\{\|\cdot\|_{(k,p)}\le\alpha\}}(w) \;\Big|\; w \in \mathbb{R}^d\Big\}, \tag{5.17}$$

where $\delta_C(\cdot)$ is the indicator function of the convex set C, and use proximal gradient methods, as discussed in earlier chapters. The problem is of the general form $\min\big\{f(w) + \lambda g(w) \;\big|\; w \in \mathbb{R}^d\big\}$, where f is a convex loss function with Lipschitz continuous gradient, and $\lambda > 0$ is a regularization parameter.

In the special case that $g(w) = \delta_C(w)$, where C is a convex set, the proximity operator reduces to the projection operator onto C, see (2.9). For the case $p = \infty$ of the (k,p)-support norm, we can compute the projection onto the unit ball using the following result.

Proposition 80. *For every $w \in \mathbb{R}^d$, the projection x of w onto the unit ball of the (k,∞)-norm is given by*

$$x_i = \begin{cases} \operatorname{sign}(w_i)(|w_i| - \beta), & \text{if } ||w_i| - \beta| \leq 1, \\ \operatorname{sign}(w_i), & \text{if } ||w_i| - \beta| > 1, \end{cases} \tag{5.18}$$

where $\beta = 0$ if $\|w\|_1 \leq k$, otherwise $\beta \in (0,\infty)$ is chosen to maximize $\sum_{i=1}^d |x_i|$ subject to the constraint $\sum_{i=1}^d |x_i| \leq k$. Furthermore, the projection can be computed in $\mathcal{O}(d \log d)$ time.

Proof. (Sketch) We assume without loss of generality that the components of w are non zero. We solve the optimization problem

$$\min_{x \in \mathbb{R}^d} \left\{ \sum_{i=1}^d (x_i - w_i)^2 \;\middle|\; |x_i| \leq 1, \sum_{i=1}^d |x_i| \leq k \right\}. \tag{5.19}$$

We consider two cases. If $\sum_{i=1}^d |w_i| \leq k$, then the problem decouples and we solve it componentwise. If $\sum_{i=1}^d |w_i| > k$, we solve problem (5.19) by minimizing the Lagrangian function $\mathcal{L}(x,\beta) = \sum_{i=1}^d (x_i - w_i)^2 + 2\beta(\sum_{i=1}^d |x_i| - k)$ with nonnegative multiplier β, which can be done componentwise. Finally, both cases can be combined into the form of (5.18). The complexity follows by taking advantage of the monotonicity of $x_i(\beta)$. □

We can use Proposition 80 to project onto the unit ball of radius $\alpha > 0$ by a rescaling argument.

Remark 81. *In order to project onto the unit ball of radius $\alpha > 0$, we solve the optimization problem $\min \left\{ \sum_{i=1}^d (x_i - w_i)^2 \;\middle|\; x \in \mathbb{R}^d,\ |x_i| \leq \alpha,\ \sum_{i=1}^d |x_i| \leq \alpha k \right\}$. To do so, we make the change of variables $x_i' = x_i/\alpha$ and note that the problem reduces to computing the projection x' of w' onto the unit ball of the norm, where $w_i' = w_i/\alpha$, which is the problem that was solved in* (5.19). *Once this is done, our solution is given by $x_i = \alpha x_i'(\beta)$, where $x'(\beta)$ is determined in accordance with Proposition 80.*

5.3.2 Matrix problem

Given data matrix $X \in \mathbb{R}^{d \times m}$ for which we observe a subset of entries, we consider the constrained optimization problem

$$\min_{W \in \mathbb{R}^{d \times m}} \left\{ \|\Omega(X) - \Omega(W)\|_{\mathrm{F}} \mid \|W\|_{(k,p)} \leq \alpha \right\} \tag{5.20}$$

where the operator Ω applied to a matrix sets unobserved values to zero. As in the vector case, the Frank-Wolfe method can be applied to the matrix problems. Algorithm 3 is particularly convenient in this case as we only need to compute the largest k singular values, which can result in a computationally efficient algorithm. The result is a direct consequence of Proposition 75 and von Neumann's trace inequality (Theorem 7). Specifically, we solve the optimization problem

$$\min_{S} \left\{ \langle S \mid G \rangle \;\middle|\; \|S\|_{(k,p)} \leq \alpha \right\},$$

where G is the gradient of the objective function at the t-th iteration. Equivalently we consider

$$\max_{R} \left\{ \langle R \mid G \rangle \;\middle|\; \|R\|_{(k,p)} \leq \alpha \right\},$$

where $R = -S$. The solution is given by

$$R = U_k \mathrm{diag}(r) V_k^{\top}$$

where U_k and V_k are the top k left and right singular vectors of the gradient G of the objective function in (5.20) evaluated at the current solution, whose singular values we denote by g, and r is obtained from g as

$$r_i = \begin{cases} \alpha \left(\frac{g_i}{\|g\|_{(k,p),*}} \right)^{\frac{1}{p-1}}, & \text{if } i \in I_k \\ 0, & \text{if } i \notin I_k \end{cases} \tag{5.21}$$

and, for $p = \infty$

$$r_i = \begin{cases} \alpha & \text{if } i \in I_k,\ g_i \neq 0, \\ 0 & \text{otherwise} \end{cases} \tag{5.22}$$

for $p \in (1, \infty)$ and $p = \infty$, respectively, where $I_k = [1, \ldots, k]$. It follows that a subgradient for the matrix problem is $R = -U_k \mathrm{diag}(s) V_k^{\top}$.

Note also that the proximity operator of the norm and the Euclidean projection on the associated unit ball both require the full singular value decomposition to be performed. Indeed, recall that the proximity operator of an orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ at $W \in \mathbb{R}^{d \times m}$ is given by $\mathrm{prox}_{\|\cdot\|}(W) = U \mathrm{diag}(\mathrm{prox}_g(\sigma(W)))V^{\top}$, where U and V are the matrices formed by the left and right singular vectors of W, see e.g. Argyriou et al. (2011, Prop. 3.1), and this requires the full decomposition.

Table 5.1: Matrix completion on synthetic datasets with decaying spectrum. The improvement of the (k,p)-support norm over the k-support norm is statistically significant at a level < 0.001.

dataset	norm	test error	k	p
rank 5	trace	0.8184 (0.03)	-	-
ρ=10%	k-supp	0.8036 (0.03)	3.6	-
	kp-supp	0.7831 (0.03)	1.8	7.3
rank 5	trace	0.4085 (0.03)	-	-
ρ=20%	k-supp	0.4031 (0.03)	3.1	-
	kp-supp	0.3996 (0.03)	2.0	4.7
rank 10	trace	0.6356 (0.03)	-	-
ρ=20%	k-supp	0.6284 (0.03)	4.4	-
	kp-supp	0.6270 (0.03)	2.0	4.4

5.4 Numerical Experiments

In this section we apply the spectral (k,p)-support norm to matrix completion (collaborative filtering problems), in which we want to recover a low rank, or approximately low rank, matrix from a small sample of its entries, see e.g. Jaggi and Sulovsky (2010). One prominent method of solving this problem is trace norm regularization: we look for a matrix which closely fits the observed entries and has a small trace norm (Jaggi and Sulovsky, 2010; Mazumder et al., 2010; Toh and Yun, 2011). In Chapter 4 we showed that the spectral k-support norm can outperform the trace norm in some applications. We now apply the (k,p)-support norm to this framework and we investigate the impact of varying p, and we compare the spectral (k,p)-support norm to the trace norm and the spectral k-support norm ($p = 2$) in both synthetic and real datasets. In each case, we solve the optimization problem (5.20) using the Frank-Wolfe method as outlined in Section 5.3. We determine the values of k and $p \geq 1$ by validation, averaged over a number of trials.

Impact of p. A key motivation for the additional parameter p is that it allows us to tune the norm to the decay of the singular values of the underlying matrix. In particular the variational formulation of (5.4) suggests that as the spectrum of the true low rank matrix flattens out, larger values of p should be preferred.

We ran the method on a set of 100×100 matrices of rank 12, with decay of the non zero singular values σ_ℓ proportional to $\exp(-\ell a)$, for 26 values of a between 10^{-6} and 0.18, and we determined the corresponding optimal value of p. Figure 5.1 illustrates the optimum value of p as a function of a. We clearly observe the negative slope, that is the steeper the slope the smaller the optimal value of p. Figure 5.2 shows the spectrum and the optimal p for several decay values.

Note that k is never equal to 1, which is a special case in which the norm is independent of p, and is equal to the trace norm. In each case the improvement of the spectral (k,p)-support norm over the k-support and trace norms is statistically significant at a level < 0.001.

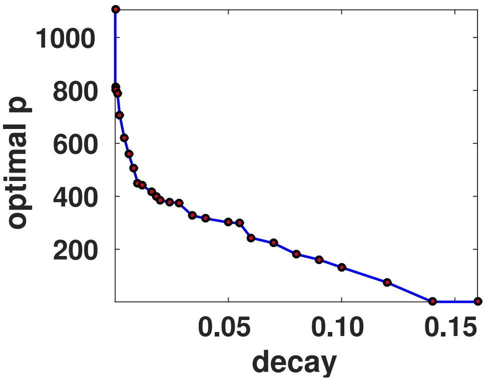


Figure 5.1: Optimal p vs. decay a, optimal k selected by validation.

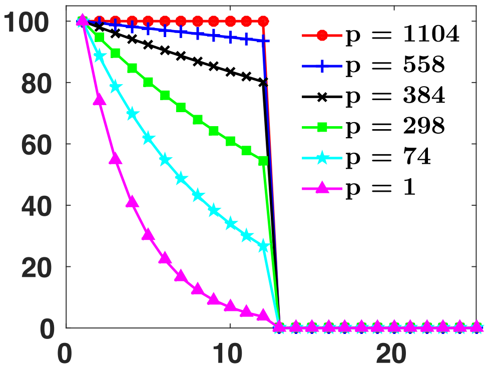


Figure 5.2: Optimal p fitted to matrix spectra with various decays.

Figure 5.3 illustrates the impact of the curvature p on the test error on synthetic and real datasets. We observe that the error levels off as p tends to infinity, so for these specific datasets the major gain is to be had for small values of p. The optimum value of p for both the real and synthetic datasets is statistically different from $p = 2$ (k-support norm), and $p = 1$ (trace norm).

Simulated Data. We replicated the experimental setting of Chapter 4 for synthetic matrix completion, see also McDonald et al. (2014b). Each 100×100 matrix is generated as $W = UV^\top + E$, where $U, V \in \mathbb{R}^{100 \times r}$, $r \ll 100$, and the entries of U, V and E are i.i.d. standard Gaussian. The error is measured as $\|\text{true} - \text{predicted}\|^2 / \|\text{true}\|^2$, standard deviations are shown in brackets and the mean values of k and p are selected by validation. Table 5.1 illustrates the results. We found that the (k,p)-support norm outperformed the standard k-support norm, as well as the trace norm, at a statistically significant level.

We note that although Frank Wolfe method for the (k,p)-support norm does not generally converge as quickly as proximal methods (which are available in the case of the k-support

Table 5.2: Matrix completion on real datasets. The improvement of the (k,p)-support norm over the k-support and trace norms is statistically significant at a level < 0.001.

dataset	norm	test error	k	p
MovieLens 100k	trace	0.2017	-	-
	k-supp	0.1990	1.9	-
	kp-supp	0.1971	2.0	3.0
Jester 1	trace	0.1752	-	-
	k-supp	0.1739	6.4	-
	kp-supp	0.1731	2.0	6.5
Jester 3	trace	0.1959	-	-
	k-supp	0.1942	2.1	-
	kp-supp	0.1924	5.0	5.0

norm as discussed in Chapter 3), the computational cost can be mitigated using a continuation method. Specifically given an ordered sequence of parameter values for p we can proceed sequentially, initializing its value based on the previously computed value. Empirically we tried this approach for a range of values of p and found that the total computation time increased only moderately.

Real Data. Finally, we applied the norms to real collaborative filtering datasets. We observe a subset of the (user, rating) entries of a matrix and predict the unobserved ratings, with the assumption that the true matrix is likely to have low rank. We report on the MovieLens 100k dataset (*http://grouplens.org/datasets/movielens/*), which consists of ratings of movies, and the Jester 1 and 3 datasets (*http://goldberg.berkeley.edu/jester-data/*), which consist of ratings of jokes. Following McDonald et al. (2014b); Toh and Yun (2011), for MovieLens for each user we uniformly sampled $\rho = 50\%$ of available entries for training, and for Jester 1 and Jester 3 we sampled 20, respectively 8 ratings per user, using 10% for validation. We used normalized mean absolute error,

$$\mathrm{NMAE} = \frac{\|\mathrm{true} - \mathrm{predicted}\|^2}{\#\mathrm{obs.}/(r_{\max} - r_{\min})},$$

where $r_{\min}$ and $r_{\max}$ are lower and upper bounds for the ratings (Toh and Yun, 2011), and we implemented a final thresholding step as in McDonald et al. (2014b).

The results are outlined in Table 5.2. The spectral (k,p)-support outperformed the trace norm and the spectral k-support norm, and the improvement is statistically significant at a level < 0.001 (the standard deviations, not shown here, are of the order of 10^{-5}). In summary, the experiments suggest that the additional flexibility of the p parameter indeed allows the model to better fit both the sparsity and the decay of the true spectrum.

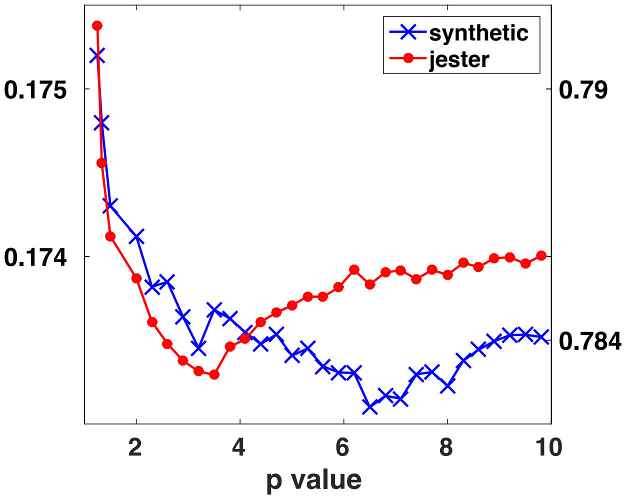


Figure 5.3: Test error vs. curvature (p). Left axis: synthetic data (blue crosses); right axis: Jester 1 dataset (red circles).

5.5 Discussion

We presented a generalization of the k-support norm, the (k,p)-support norm, where the additional parameter p is used to better fit the decay of the components of the underlying model. The norm belongs to the Θ p-norm family introduced in Chapter 3. We studied its properties and found that while the additional parameter requires us to validate over a larger space, numerical experiments in low rank matrix completion justify the additional cost.

Chapter 6

The Tensor k-Support Norm

In previous chapters we considered the well known problems of learning a sparse vector or a low rank matrix. A related problem which increasingly is being explored in recent years is learning a tensor from a set of linear measurements. Applications include collaborative filtering (Karatzoglou et al., 2010), medical imaging (Gandy et al., 2011), multitask learning (Romera-Paredes et al., 2013), video processing (Liu et al., 2009), and many more, see e.g. the recent review by Cichocki et al. (2015) and references therein.

The number of entries of an N-way tensor grows exponentially with the number of modes, hence it is crucial to impose prior knowledge into the learning problem in order to avoid overfitting. In line with the approaches that we have investigated in previous chapters, a common approach is to encourage low rank tensors via convex penalties which approximate the rank function. However, unlike the matrix case, there are several notions of tensor rank that one may consider (see e.g. Kolda and Bader, 2009, for a review) and it is not apparent which one is most effective for a particular application. A natural notion of tensor rank is that of CP-rank: the minimum number of rank one tensors, each formed by the outer product of N vectors of the corresponding dimensions, which sum to the original tensor. In the case $m = 2$ the CP-rank reduces to the standard rank of a matrix. Unfortunately computing the CP-rank is an NP-hard problem (Hillar and Lim, 2013). A more tractable alternative, which has been considered in a number of papers, is provided by the Tucker rank (Gandy et al., 2011; Liu et al., 2009; Romera-Paredes and Pontil, 2013; Signoretto et al., 2014). This is the vector composed by the rank of every matricization of a tensor.

Even though the Tucker rank can be computed efficiently (for example via the singular value decomposition of each matricization), deriving a convex relaxation remains difficult. By convex relaxation we mean a convex function, typically a norm, whose unit ball approximates well the set of low rank tensors over a prescribed bounded convex set. An alternative approach considers the convex envelope (Rockafellar, 1970) of the sum of the components of the Tucker rank over a bounded convex set. In the latter setting Gandy et al. (2011)

proposed a relaxation formed by the sum of the nuclear norms of each matricization of a tensor. However, as argued by Romera-Paredes and Pontil (2013) this relaxation is not tight. Their improved relaxation nevertheless is difficult to employ within first order optimization algorithms, because the proximity operator cannot be computed in a finite number of steps. Yet another approach is provided by Mu et al. (2014), however it can be applied only with tensors of four or more modes.

In this chapter, we propose a tight convex relaxation for the Tucker rank which is linked to the notion of the k-support norm. We consider the set defined by the convex hull of the union of tensors whose matricizations have bounded rank.

We show that the convex hull of this set is the unit ball of a norm which we call the tensor k-support norm, and provide the expression for the dual norm. A special case of this formulation yields the latent tensor trace norm of Tomioka et al. (2011).

For simplicity we focus on the problem of tensor completion, that is, learning a tensor from a subset of its entries. We present numerical experiments on one synthetic dataset and three real datasets, which indicate that the proposed method improves significantly over the method by Tomioka et al. (2011) (see also Tomioka and Suzuki, 2013) in terms of estimation error, while remaining computationally efficient.

The chapter is organized as follows. In Section 6.1 we review basic concepts and terminology for tensors. In Section 6.2 we present the tensor k-support norm and we establish some of its basic properties. In Section 6.3 we discuss a method to solve the corresponding regularization problem. In Section 6.4 we present our numerical experiments. Section 6.5 contains final remarks.

6.1 Background

In this section we recall the notion of the k-support norm and introduce some basic terminology from multilinear algebra. Recall from Chapter 5 that for every $k \in \{1, \ldots, d\}$, and for every $p \in [1, \infty]$, the (k,p)-support norm $\|\cdot\|_{(k,p)}$ is defined as the norm whose unit ball is given by

$$\mathrm{conv}\left\{w \in \mathbb{R}^d \;\middle|\; \mathrm{card}(w) \le k,\ \|w\|_p \le 1\right\},$$

that is, the convex hull of the set of vectors of cardinality at most k and ℓ_p-norm no greater than one. For $p = 2$ the norm is the standard k-support norm of Argyriou et al. (2012), and we denote it by $\|\cdot\|_{(k)}$. Recall further that the dual of the (k,p)-support norm is given, for every $u \in \mathbb{R}^d$, by

$$\|u\|_{(k,p),*} = \left(\sum_{i=1}^{k} (|u|_i^{\downarrow})^q\right)^{\frac{1}{q}} \tag{6.1}$$

where $|u|^{\downarrow}$ is the vector obtained from u by reordering its components so that they are non-increasing in absolute value, and $q \in [1, \infty]$ is computed by the formula $\frac{1}{p} + \frac{1}{q} = 1$.

6.1.1 Multilinear algebra

We now introduce some notions from multilinear algebra. Let N be an integer greater than one and let $d_1, \ldots, d_N$ be strictly positive integers. Let $\boldsymbol{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ be an N-way tensor, that is

$$\boldsymbol{W} = \big(W_{i_1, \ldots, i_N} \bigm| 1 \leq i_1 \leq d_1, \ldots, 1 \leq i_N \leq d_N\big).$$

A mode-n fiber is a vector formed by the elements of $\boldsymbol{W}$ obtained by fixing all indices but those corresponding to the n-th mode. For every $n \in \{1, \ldots, N\}$, let $D_n = \prod_{\ell \neq n} d_\ell$. The n-th matricization of a tensor $\boldsymbol{W}$, denoted by $W^{(n)}$, is the $d_n \times D_n$ matrix obtained by arranging the mode-n fibers of $\boldsymbol{W}$ such that each of them forms a column of $\boldsymbol{W}^{(n)}$. By way of example, a 3-way tensor $\boldsymbol{W} \in \mathbb{R}^{3 \times 4 \times 2}$ admits the matricizations $W^{(1)} \in \mathbb{R}^{3 \times 8}$, $W^{(2)} \in \mathbb{R}^{4 \times 6}$, and $W^{(3)} \in \mathbb{R}^{2 \times 12}$. We also denote by $M_n : \mathbb{R}^{d_1 \times \cdots \times d_N} \to \mathbb{R}^{d_n \times D_n}$ the linear transformation which associates a tensor to its n-th matricization, that is, $M_n(\boldsymbol{W}) = W^{(n)}$. Note that the adjoint of M_n, that is, the operator $M_n^\top : \mathbb{R}^{d_n \times D_n} \to \mathbb{R}^{d_1 \times \cdots \times d_N}$, is the reverse matricization along mode n.

In contrast to the matrix setting, the rank of a tensor can be defined in a number of ways. Two important notions are the CP-rank and Tucker-rank. The CP-rank of an N-way tensor $\boldsymbol{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is defined as the smallest nonnegative R such that

$$\boldsymbol{W} = \sum_{r=1}^{R} a_1^{(r)} \otimes \cdots \otimes a_N^{(r)} \tag{6.2}$$

for some vectors $a_n^{(r)} \in \mathbb{R}^{d_n}$, $n \in \{1, \ldots, N\}$, $r \in \{1, \ldots, R\}$. In other words, the CP-rank is the minimum number of rank-one tensors which sum to $\boldsymbol{W}$, where each rank-one tensor is an outer product of N vectors of the corresponding dimensions. Computing the CP-rank is an NP-hard problem (Hillar and Lim, 2013), hence it is a difficult notion to which to apply a convex relaxation in order to lead to convex regularization problems. In fact, simply computing the nuclear norm of a tensor is an NP-hard problem (Friedland and Lim, 2016). This has motivated the study of simpler notions of tensor rank, which can be computed efficiently and which are amenable to tractable convex relaxations. In this work we follow a recent line of works (Gandy et al., 2011; Liu et al., 2009; Romera-Paredes and Pontil, 2013; Signoretto et al., 2014) which employ the notion of Tucker rank. This is defined as the vector formed by the rank of each matricization of a tensor, that is

$$\text{Tucker-rank}(\boldsymbol{W}) = \big(\text{rank}(W^{(1)}),\ldots,\text{rank}(W^{(N)})\big).$$

In the vector case, the convex relaxation of the cardinality operator, also referred to as the ℓ_0-'norm', on the unit ℓ_∞-ball is the ℓ_1-norm (Fazel et al., 2001) as discussed in Chapter 2, and likewise the convex envelope of the rank function on the unit spectral-norm ball is the nuclear norm[1].

In the multilinear setting, a popular approach to learn low rank tensors is to apply low rank matrix regularizers to the matricizations of a tensor and a number of these have been studied. The *overlapped nuclear norm*[2] is defined (Gandy et al., 2011; Liu et al., 2009; Romera-Paredes and Pontil, 2013; Tomioka et al., 2011) as the sum of the nuclear norms of the mode-n matricizations, namely

$$\|\boldsymbol{W}\|_{\text{over}} = \sum_{n=1}^{N} \|M_n(\boldsymbol{W})\|_{\text{nuc}}.$$

The corresponding dual norm of a tensor $\boldsymbol{U}$ is

$$\|\boldsymbol{U}\|_{\text{over},*} = \inf\left\{ \max_{1\le n\le N} \|M_n(\boldsymbol{V}_n)\|_{\text{sp}} \;\Big|\; \sum_{n=1}^{N} \boldsymbol{V}_n = \boldsymbol{U} \right\}.$$

Tomioka and Suzuki (2013) observed that the overlapped nuclear norm does not perform well when a tensor is only low-rank in one of its modes. This motivated the authors to consider a different norm. Specifically, let $\alpha_1,\ldots,\alpha_N$ be strictly positive reals. The *scaled latent nuclear norm* is defined by the variational problem

$$\|\boldsymbol{W}\|_{\text{lat}} = \inf\left\{ \sum_{n=1}^{N} \frac{1}{\alpha_n} \|M_n(\boldsymbol{V}_n)\|_{\text{nuc}} \;\Big|\; \sum_{n=1}^{N} \boldsymbol{V}_n = \boldsymbol{W} \right\}. \tag{6.3}$$

The case $\alpha_n = 1$ is presented in Tomioka and Suzuki (2013), and we will refer to this as the latent tensor trace norm. The case $\alpha_n = \sqrt{d_n}$ is presented in Wimalawarne et al. (2014). We also observe that the dual norm of (6.3) is given, for every tensor $\boldsymbol{U} \in \mathbb{R}^{d_1\times\cdots\times d_N}$ by the expression

$$\|\boldsymbol{U}\|_{\text{lat},*} = \max_{1\le n\le N} \alpha_n \|M_n(\boldsymbol{U})\|_{\text{sp}}.$$

We next present an extension of this formulation which is linked to the k-support norm, and which has a natural interpretation as a convex relaxation of a set of tensors with bounded

[1] For consistency with the tensor literature, we refer to the norm as the *nuclear* norm rather than trace norm throughout this chapter.

[2] Also sometimes referred to as *sum of nuclear norms* in Gandy et al. (2011); Mu et al. (2014) and others.

Tucker rank.

6.2 Tensor k-Support Norm

In this section, we introduce the proposed norm and establish some of its basic properties. Our starting point is to consider the set of tensors for which the rank of every matricization is bounded by some value. Specifically, for each mode $n \in \{1, \dots, N\}$ we choose a positive integer $k_n \in \{1, \dots, d_n\}$ which bounds the rank of the n-th matricization, a real $p_n \in [1, \infty]$ and we define the set

$$\mathcal{A}_n = \Big\{ \boldsymbol{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N} \;\Big|\; \operatorname{rank}(W^{(n)}) \leq k_n, \|W^{(n)}\|_{p_n} \leq 1 \Big\} \tag{6.4}$$

where $\|\cdot\|_{p_n}$ is the Schatten p_n-norm of a matrix. A simpler setting is $k_n = k$ and $p_n = 2$ for all $n \in \{1, \dots, N\}$, however for now we allow for the most general case. The sets $\mathcal{A}_n$ can be used to construct a unit ball of a norm in two natural ways, either by considering their union or their intersection. The latter leads to a problem which is not tractable, and which requires approximating, hence we leave the analysis to a later date. Following the former approach, we consider the set of tensors

$$\mathcal{C} = \operatorname{conv} \bigcup_n \mathcal{A}_n, \tag{6.5}$$

that is the convex hull of the set of tensors for which at least one matricization is bounded in norm and in rank in accordance with the constraints in (6.4). The set $\mathcal{C}$ is bounded, convex, balanced, and absorbing, hence by Lemma 36 it is the unit ball of a norm, which we denote by $|||\cdot|||$. This norm is given by the Minkowski functional of the set $\mathcal{C}$, that is

$$|||\boldsymbol{W}||| \;= \inf \Big\{ \lambda \;\Big|\; \lambda > 0, \frac{1}{\lambda} \boldsymbol{W} \in \mathcal{C} \Big\}. \tag{6.6}$$

Note that the alternative of considering $\operatorname{conv} \bigcap_n \mathcal{A}_n$ corresponds to the set of tensors for which each matricization is bounded in norm and in rank. We do not expand further on this construction in this work.

The next result characterizes the norm and the associated dual norm.

Proposition 82. *For every tensor $\boldsymbol{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ we have that*

$$|||\boldsymbol{W}||| = \inf \left\{ \sum_{n=1}^{N} \|M_n(\boldsymbol{V}_n)\|_{(k_n, p_n)} \;\Big|\; \sum_{n=1}^{N} \boldsymbol{V}_n = \boldsymbol{W} \right\} \tag{6.7}$$

where the infimum is over the tensors $\boldsymbol{V}_n \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ for $n \in \{1, \dots, N\}$.

Furthermore the dual norm of $\boldsymbol{U} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$ is given by

$$|||\boldsymbol{U}|||_* = \max_{1 \le n \le N} \|M_n(\boldsymbol{U})\|_{(k_n, p_n),*}$$

Proof. To simplify our presentation we use the shorthand notation $\|\boldsymbol{W}\|_{[n]} = \|M_n(\boldsymbol{W})\|_{(k_n,p_n)}$ and $\|\boldsymbol{U}\|_{[n],*} = \|M_n(\boldsymbol{U})\|_{(k_n,p_n),*}$. The expression for the primal norm follows by noting that

$$\mathcal{C} = \operatorname{conv} \bigcup_n \left\{ \boldsymbol{W} \in \mathbb{R}^{d_1 \times \cdots \times d_N} \;\middle|\; \|\boldsymbol{W}\|_{[n]} \le 1 \right\}$$

and using the formula (6.6). To identify the dual norm, we note, for every $\boldsymbol{U} \in \mathbb{R}^{d_1 \times \cdots \times d_N}$, that

$$\begin{aligned}
|||\boldsymbol{U}|||_* &= \sup \left\{ \langle \boldsymbol{U} \mid \boldsymbol{W} \rangle \;\middle|\; |||\boldsymbol{W}||| \le 1 \right\} \\
&= \sup \left\{ \langle \boldsymbol{U} \mid \boldsymbol{W} \rangle \;\middle|\; \boldsymbol{W} \in \mathcal{A}_n,\ 1 \le n \le N \right\} \\
&= \sup \left\{ \max_{1 \le n \le N} \left\langle U^{(n)} \mid W^{(n)} \right\rangle \;\middle|\; \boldsymbol{W} \in \mathcal{A}_n \right\} \\
&= \max_{1 \le n \le N} \sup \left\{ \left\langle U^{(n)} \mid W^{(n)} \right\rangle \;\middle|\; \boldsymbol{W} \in \mathcal{A}_n \right\} \\
&= \max_{1 \le n \le N} \|U^{(n)}\|_{(k_n, p_n),*}
\end{aligned}$$

where, for each $n \in \{1, \dots, N\}$, $U^{(n)} = M_n(\boldsymbol{U})$, $W^{(n)} = M_n(\boldsymbol{W})$ and we have used the fact that $\langle \boldsymbol{W} \mid \boldsymbol{U} \rangle = \left\langle W^{(n)} \mid U^{(n)} \right\rangle$ for every n and the fact that the maximum of a linear functional over a compact convex set is attained at an extreme point of that set (Proposition 42). The last equality follows by Equation (6.1). □

At last, we note that setting $k_n = 1$ and $p_n = 1$, for every $n \in \{1, \dots, N\}$ we recover the latent tensor trace norm presented in Tomioka and Suzuki (2013), cf. Equations (6) and (7).

6.2.1 Rademacher bound

We now give a bound on the empirical Rademacher average of the linear function class associated with the tensor k-support norm. This quantity plays a key role in deriving uniform bounds in empirical risk minimization, see for example Bartlett and Mendelson (2002).

Let m be a strictly positive integer. Given a set of data tensors $\mathbf{X}_1, \cdots, \mathbf{X}_m \in \mathbb{R}^{d_1 \times \ldots \times d_N}$, the Rademacher average of the class $\left\{ \mathbf{X} \mapsto \langle \boldsymbol{W} \mid \mathbf{X} \rangle \;\middle|\; |||\boldsymbol{W}||| \le 1 \right\}$ is defined as

$$\mathcal{R} = \frac{2}{n} \mathbb{E}_\epsilon \sup_{|||\boldsymbol{W}||| \le 1} \sum_{j=1}^m \epsilon_j \langle \mathbf{X}_j \mid \boldsymbol{W} \rangle$$

where $\epsilon_1, \dots, \epsilon_n$ are i.i.d. Rademacher random variables. Note that the above expression depends on the data tensors. In particular, in the case of tensor completion the data tensors are masks, each of which measures one entry in the tensor. If Ω is the set of observed entries, the corresponding empirical Rademacher average, which we denote by $\mathcal{R}_\Omega$, has the form

$$\mathcal{R}_\Omega = \frac{2}{n}\mathbb{E}_\epsilon \sup_{|||\boldsymbol{W}|||\leq 1} \sum_{(i_1,\ldots,i_N)\in\Omega} \epsilon_{i_1,\ldots,i_N} W_{i_1,\ldots,i_N}.$$

Proposition 83. *There holds the bound*

$$\mathcal{R} \leq \frac{2}{m} \max_{1\leq n\leq N} k_n^{\frac{1}{q_n}} \mathbb{E}_\epsilon \left\| M_n\Big(\sum_{j=1}^m \epsilon_j \boldsymbol{X}_j\Big)\right\|_{\mathrm{sp}} + 8C\sqrt{\frac{\log N}{m}}$$

where $q_n \in [1,\infty]$ *satisfies* $\frac{1}{p_n}+\frac{1}{q_n}=1$ *and*

$$C = \sqrt{\max_{n=1}^{N} \sup_{\|M_n(\boldsymbol{W})\|_{(k_n,p_n)}\leq 1} \frac{1}{m}\sum_{j=1}^m \langle \boldsymbol{W} \,|\, \boldsymbol{X}_j\rangle^2}.$$

In particular for tensor completion we have

$$\mathcal{R}_\Omega \leq 2k^*\left[\frac{1}{\sqrt{m}} + \sqrt{\frac{2\log(N+1)}{m}}\right] + 8\frac{\sqrt{\log N}}{m}$$

where $k^* = \max_{1\leq n\leq N} k_n^{\frac{1}{q_n}}$.

Proof. See Appendix E. □

An interesting feature of the bound is that it depends only logarithmically on the number of modes N, and it depends only on the maximum upper bound on the rank of the matricizations. Using this bound one may then apply the well established technique in Bartlett and Mendelson (2002) to derive associated risk bounds.

6.3 Optimization

In this section, we consider regularization with the tensor k-support norm. We restrict our analysis to the case $p_n = 2$ for all $n \in \{1,\ldots,N\}$ as the proximity operator for the corresponding vector norm is known (Chatterjee et al., 2014), but to our knowledge it is not known in the general case $p_n \neq 2$. Furthermore, to ease our presentation we restrict our discussion to the setting of tensor completion, however our observations can be easily extended to the more general problem of estimating a tensor from a set of linear measurements; an important such case which may be studied in a future work is multilinear multitask learning (Romera-Paredes and Pontil, 2013).

Let $\boldsymbol{W} \in \mathbb{R}^{d_1\times\ldots\times d_N}$ be an underlying (unknown) tensor we wish to learn. We consider the linear model

$$y = \mathcal{L}(\boldsymbol{W}) + \xi$$

where $\mathcal{L} : \mathbb{R}^{d_1 \times \ldots \times d_N} \to \mathbb{R}^m$ is a sampling operator which measures m linear functionals and the components of the vector $\xi \in \mathbb{R}^m$ are i.i.d. zero mean noise variables, e.g. zero mean Gaussians. In tensor completion the linear operator $\mathcal{L}$ samples a subset

$$\Omega \subset \{1, \ldots, d_1\} \times \cdots \times \{1, \ldots, d_N\}$$

of the tensor entries[3]. We also let $\boldsymbol{Y} \in \mathbb{R}^{d_1 \times \ldots \times d_N}$ be any tensor such that $\mathcal{L}(\boldsymbol{Y}) = y$.

We denote by $\mathcal{P}_\Omega$ the orthogonal projector onto the linear space formed by all tensors, the components of which are zero outside the index set Ω. That is, for every $\boldsymbol{W} \in \mathbb{R}^{d_1 \times \ldots \times d_N}$ and

$$(i_1, \ldots, i_N) \in \{1, ..., d_1\} \times \cdots \times \{1, ..., d_N\},$$

the $(i_1, \ldots, i_N)$-component of $\mathcal{P}_\Omega(\boldsymbol{W})$ equals to $W_{i_1, \ldots, i_N}$ if $(i_1, \ldots, i_N) \in \Omega$ and zero otherwise.

We attempt to estimate the tensor by solving the Tikhonov regularization problem

$$\underset{\boldsymbol{W}}{\text{minimize}} \ \frac{1}{2} \|\mathcal{P}_\Omega(\boldsymbol{W}) - \mathcal{P}_\Omega(\boldsymbol{Y})\|_2^2 + \lambda |||\boldsymbol{W}|||, \tag{6.8}$$

where the norm $|||\cdot|||$ is given by equation (6.7) and λ is a positive parameter which may be chosen by cross-validation. We see that problem (6.8) is equivalent to the problem of minimizing the objective function

$$\frac{1}{2} \left\| \mathcal{P}_\Omega \Big(\sum_{n=1}^{N} \boldsymbol{V}_n - \boldsymbol{Y} \Big) \right\|_2^2 + \lambda \sum_{n=1}^{N} \|M_n(\boldsymbol{V}_n)\|_{(k_n)}, \tag{6.9}$$

over the tensors $\boldsymbol{V}_n \in \mathbb{R}^{d_1 \times \ldots \times d_N}$, $n \in \{1, \ldots, N\}$.

This optimization problem is of the general form

$$\underset{x \in \mathcal{H}}{\text{minimize}} f(x) + g(x) \tag{6.10}$$

where $x \equiv (\boldsymbol{V}_1, \cdots, \boldsymbol{V}_N)$ and $\mathcal{H}$ is a Hilbert space (in our case finite dimensional), and can be solved by proximal gradient methods as discussed in previous chapters. In particular, we employ the forward-backward algorithm (Combettes and Pesquet, 2011), which consists of

[3] For simplicity we assume that each entry is sampled at most once.

the iterative scheme

$$x_t = \operatorname{prox}_{sg}(x_{t-1} - s\nabla f(x_{t-1})), \quad t \in \mathbb{N} \tag{6.11}$$

for some starting point $x_0 \in \mathcal{H}$ and step size $s > 0$. It is well-known that the iterates (6.11) converge to a solution of the optimization problem (6.10) provided the function f is smooth with L-Lipschitz continuous gradient and $s \leq 2/L$. Furthermore, for such values of s, the sequence of objective values $\big\{f(x_t) + g(x_t) \bigm| t \in \mathbb{N}\big\}$ is monotonically non increasing (see e.g. Combettes and Pesquet, 2011, for a review).

In our setting we need to solve the optimization problem

$$\underset{\boldsymbol{V}_1,\ldots,\boldsymbol{V}_M}{\text{Minimize}}\ f(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N) + g(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N)$$

where we defined

$$f(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N) = \frac{1}{2}\Big\|\mathcal{P}_\Omega\Big(\sum_{n=1}^{N}\boldsymbol{V}_n - \boldsymbol{Y}\Big)\Big\|_2^2$$

and

$$g(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N) = \lambda\sum_{n=1}^{N}\|M_n^\top \boldsymbol{V}_n\|_{(k_n)}.$$

Clearly the error term f is smooth. The gradient of $\frac{1}{2}||\mathcal{P}_\Omega(\boldsymbol{W} - \boldsymbol{Y})||^2$ w.r.t. $\boldsymbol{W}$ is simply given by the tensor $\mathcal{P}_\Omega(\boldsymbol{W} - \boldsymbol{Y})$, that is the tensor, the entries of which are equal to that of $\boldsymbol{W} - \boldsymbol{Y}$ if the entry belongs to the set Ω, and zero otherwise. Using this observation we obtain, for every $n \in \{1,\ldots,N\}$, the formula

$$\nabla_{\boldsymbol{V}_n} f(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N) = \mathcal{P}_\Omega\Big(\sum_{n=1}^{N}(\boldsymbol{V}_n - \boldsymbol{Y})\Big).$$

Note also that since $\mathcal{P}_\Omega$ is an orthogonal projection, the Lipschitz constant of the gradient is equal to 1.

Next we observe that the proximity operator of the function g at $(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N)$ is given by the concatenation of the proximity operators of the functions $\lambda\|\cdot\|_{(k_1)},\ldots,\lambda\|\cdot\|_{(k_N)}$, that is

$$\operatorname{prox}_g(\boldsymbol{V}_1,\ldots,\boldsymbol{V}_N) = \big(\operatorname{prox}_{\lambda\|\cdot\|_{(k_n)}}(\boldsymbol{V}_n)\big)_{1\leq n\leq N}.$$

Finally, given the proximity operator of a symmetric gauge function g, using von Neumann's trace inequality (Von Neumann, 1937) the proximity operator of its induced orthogonally invariant norm $\|\cdot\| = g(\sigma(\cdot))$ is given, for every $W^{(n)} \in \mathbb{R}^{d_n\times D_n}$, by the formula

$$\operatorname{prox}_{s\|\cdot\|}(W^{(n)}) = U\operatorname{diag}(\operatorname{prox}_g(\sigma(W^{(n)})))V^\top$$

where U and V are the matrices formed by the left and right singular vectors of $W^{(n)}$. It follows that we can compute the proximity operator each norm $\|\cdot\|_{k_n}$ using the formula for the proximity operator of the vector norm, which is detailed in Chatterjee et al. (Theorem 2 2014).

At last we note that when the squared error term is replaced by a smooth function we still have a simple expression of the gradient, and the above method can be readily applied. On the other hand, when the error function is convex but not smooth we may solve the regularization problem by the Douglas-Rachford optimization method (see e.g. Combettes and Pesquet, 2011).

6.4 Experiments

In this section we report on the performance of the tensor k-support norm (*ks*) in low rank tensor completion experiments. We consider the case $p_n = 2$ for all n and compare a range of settings for $(k_1, \ldots, k_N)$ to the latent tensor trace norm (*tr*) (Tomioka et al., 2011) which corresponds to the case $k_n = 1$ for all $n \in \{1, \ldots, N\}$. For the optimization we used the proximal gradient method outlined in Section 6.3. As convergence criterion we use a relative tolerance of 10^{-3} in the iterates, we average each experiment over a number of trials and we perform a *t*-test at a level $p < 0.001$ to confirm statistical significance of the improvements. Our goal with the experiments is to show that by tuning one or more of the k_n parameters we can improve learning relative to the trace norm. While the potential number of combinations of parameters $(k_1, \ldots, k_N)$ is large, our aim is not to exhaustively explore the space, but to demonstrate that for some reasonable ranges we can obtain statistically significant improvement, without the brute force approach.

6.4.1 Simulated data

Each $d \times d \times d$ tensor was generated with CP-rank R as $\boldsymbol{W} = A \circ B \circ C + \xi$, where each of A, B and C is $d \times R$, with i.i.d. standard Gaussian entries, and ξ is a tensor of i.i.d. noise with a specified standard deviation.

We set $d = 40$, $R \in \{5, 10\}$ and for simplicity we considered the parameter settings $k_1 = k_2 = k_3 = k$, for $k \in \{1, 2, \ldots, d\}$. We choose k by validation, along with the regularization parameter $\lambda \in 10^{(-3, -2.75, \ldots, 3)}$. We sampled uniformly a percentage $\rho \in \{5\%, 10\%, 20\%\}$ of the observed entries for training, used 10% for validation, and allocated the remaining entries for testing. The errors were measured as $\|\text{true} - \text{predicted}\| / \|\text{true}\|$ following Tomioka et al. (2011), and we report the mean test error and standard deviation over 100 trials. Table 6.1 illustrates the results.

We note that the tensor k-support norm outperforms the latent trace norm and the improvement is statistically significant. For each trial the optimal value of k identified by

Table 6.1: Tensor completion on synthetic datasets. $\boldsymbol{W} \in \mathbb{R}^{40\times 40\times 40}$, rank $r \in \{5, 10\}$, training set size $\rho \in \{5\%, 10\%, 20\%\}$. Test error and mean optimal k are shown. The improvement of *ks* relative to *tr* is significant ($p < 0.001$).

Data	**Norm**	**Test Err (STD)**	k
rank 5	*tr*	0.9300 (0.0103)	-
$\rho = 5\%$	*ks*	0.9297 (0.0105)	2.29
rank 5	*tr*	0.8228 (0.0172)	-
$\rho = 10\%$	*ks*	0.8217 (0.0170)	1.91
rank 5	*tr*	0.6892 (0.0235)	-
$\rho = 20\%$	*ks*	0.6891 (0.0235)	2.25
rank 10	*tr*	0.9603 (0.0047)	-
$\rho = 5\%$	*ks*	0.9601 (0.0048)	2.34
rank 10	*tr*	0.8735 (0.0088)	-
$\rho = 10\%$	*ks*	0.8728 (0.0087)	2.97
rank 10	*tr*	0.7319 (0.0130)	-
$\rho = 20\%$	*ks*	0.7316 (0.0129)	3.85

validation is an integer, and we report their mean value.

As shown by Tomioka et al. (2011), the latent tensor trace norm outperforms the tensor trace norm (Gandy et al., 2011; Liu et al., 2009), that is the sum of nuclear norms of the matricizations of a tensor, in a number of tensor completion settings, so we do not compare to this method here. We further compared the results to the Euclidean, or ℓ_2-norm, but the performance of the latter was poor, and we do not report on it here.

6.4.2 Real datasets

Schools Dataset. We next report our empirical results on the Inner London Education Authority (ILEA) dataset, *Schools* (http://www.bristol.ac.uk/cmm/learning/support/datasets). This comprises examination marks between 1 and 70 of 15362 students each of whom is described by categorical attributes. Following Romera-Paredes and Pontil (2013) we consider five attributes, corresponding to school, ethnicity, year, gender and verbal reasoning band. This gives rise to a sparse 5-way tensor $\boldsymbol{W} \in \mathbb{R}^{139\times 11\times 3\times 2\times 3}$ whose values are the mean examination mark of all students with the corresponding attributes, with a total of 3998 entries.

We uniformly sampled a training set of size $\rho = 30\%$. We used 10% of the entries for validation, and used the remaining entries for testing. As in the vector case, an advantage of the tensor k-support norm is that tailoring the k parameters can allow the model to better fit the data, at the computational cost of additional validation efforts. When the modes of the data tensor correspond to inhomogeneous attributes, varying the corresponding k_n

Table 6.2: Tensor completion on real datasets. School exams ($139 \times 11 \times 3 \times 2 \times 3$), Knee MRI ($256 \times 256 \times 27$) and Brain MRI ($256 \times 256 \times 22$) datasets. Training size $\rho = 30\%$, $(k, 1, 1, 1, 1)$, $k \in \{1, 2, \ldots, 30\}$ for Schools, and $(k, k, 1)$, $k \in \{1, 3, 6, 9, 15, 18\}$ for MRI datasets.

Data	**Norm**	**Test Err (STD)**	$(k_1, \ldots, k_n)$
Schools	*tr*	22.682 (0.1785)	-
$\rho = 30\%$	*ks*	22.464 (0.1851)	$(12, 1, 1, 1, 1)$
Knee	*tr*	0.7606 (7.53e-4)	-
$\rho = 30\%$	*ks*	0.5245 (7.32e-4)	$(6, 6, 1)$
Brain	*tr*	0.7920 (9.33e-4)	-
$\rho = 30\%$	*ks*	0.5961 (4.69e-4)	$(12, 12, 1)$

parameters can be done in a targeted manner rather than brute force testing all combinations. In this instance the *school* category is by some margin the largest dimension. Moreover, recall that along mode n we require $k_n \leq d_n$. Consequently we considered $(k, 1, 1, 1, 1)$, for $k \in \{1, 2, \ldots, 30\}$, in addition to $(2, 2, 2, 2, 2)$ and $(139, 11, 3, 2, 3)$, where the latter corresponds to the Euclidean norm. This gave a total of 32 parameter settings which we explored via validation. Following Romera-Paredes and Pontil (2013) we measured the error as root mean square error (RMSE).

The results are presented in Table 6.2. The improvement over the latent tensor trace norm is statistically significant at a level $p < 0.001$. The results suggest that tuning selected k_n can lead to improved performance over the latent tensor trace norm.

MRI. The second set of datasets comes from OsiriX (http://www.osirix-viewer.com/resources/dicom-image-library). These are collections of MRI scans of patients. Each collection is a series of cross sectional images of size 256×256, which when stacked correspond to a three dimensional representation of the underlying body part. We analyzed two datasets: *Knee*, of size $256 \times 256 \times 27$, and *Brain*, of size $256 \times 256 \times 22$, which we then standardized. Compared to the Schools dataset, the tensors are an order of magnitude larger, hence choosing a reasonably small number of parameter settings is important to minimize additional computational overhead. To account for the relative sizes of the dimensions, we considered $(k, k, 1)$ for $k \in \{1, 3, 6, 9, 12, 15\}$ for both datasets. We used 30% for training, 10% for validation, and ran 10 trials with RMSE as before. The results are outlined in Table 6.2. The k-support norm outperforms the baseline latent tensor trace norm, even for this modest selection of k values, with statistical significance at a level $p < 0.001$. In both cases the optimal setting was for $k \neq 1$. With additional tuning over a wider selection of parameters the error could reasonably be further reduced, however as our goal is to motivate using k-support norm we do leave extensive computer simulations to another occasion.

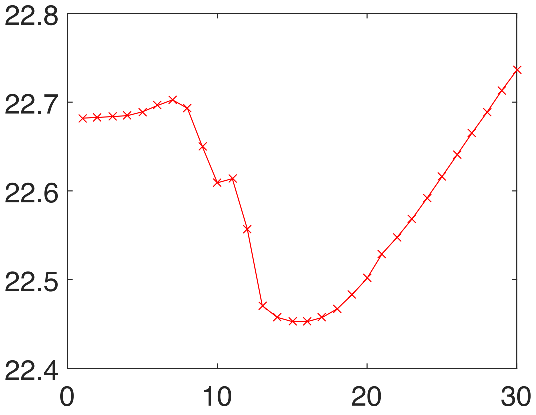


Figure 6.1: Test error vs. k for $(k,1,1,1,1)$, $k \in \{1,2,\ldots,30\}$ on Schools dataset $(139 \times 11 \times 3 \times 2 \times 3)$. Improvement vs. the latent tensor trace norm $(k=1)$ without exhaustively exploring the entire parameter space.

Parameter Selection. Our experimental results show that the additional hyperparameters k_n can lead to improved performance relative to the latent tensor trace norm, however it is necessary to be selective in the choice of parameter ranges. With real datasets in particular, we would typically expect the modes to be inhomogeneous, and potentially with large differences in dimension. Our results suggest that larger k values may be preferred along modes where the dimension is high, and smaller k values for the lower dimensional modes. In such cases our experiments suggest that by tuning k_n only along the larger dimension(s), and setting the remaining k_n values to 1, performance can be improved relative to the latent tensor trace norm, without fully exploring the potential parameter space. Figure 6.1 illustrates this effect on the Schools dataset, and shows the test error for $(k,1,1,1,1)$, with $k \in \{1,2,\ldots,30\}$. The x-axis corresponds to the value of k, with the test error for the latent tensor trace norm on the far left $(k=1)$, and we found that performance depends reasonably on the value of the parameter k_1, hence by tuning this parameter only we can obtain improved performance, without exhaustively exploring the entire potential parameter space.

6.5 Discussion

In this chapter we presented the tensor k-support norm for learning low Tucker rank tensors. In its simplest form, the unit ball of the norm is the convex hull of the union of sets of tensors, for which one of the matricizations has rank bounded by k and Frobenius norm bounded by one. We established some basic properties of this norm and addressed numerical algorithms for solving the associated regularization problem. Our numerical experiments indicate that the norm performs well and is a valuable alternative to other convex regularizers for learning tensors.

In future work, we would like to further study risk bounds for empirical error minimization with the proposed norm. On the optimization side, it would be interesting to study iterative schemes to solve the regularization problem in the general case that $p \neq 2$. This may be possible for Ivanov regularization when $p = \infty$, following ideas in McDonald et al. (2016a). A further area of study is the case where we consider as the unit ball the set comprised of the intersection of the sets $\mathcal{A}_n$, which leads to an optimization problem involving the proximity operator of the (k,p)-support norm, which, as we mentioned in Chapter 5, is not known for finite values of $p \neq 2$.

At last we note that the proposed norm naturally extends to the case in which we wish to bound the rank of generalized matricizations of a tensor, which are associated with a partition of the set of modes $\{1, \ldots, N\}$ into two sets. When all such partitions are taken into account, this construction would give rise to a norm which forms a relaxation for the so-called matrix-product-state decompositions (Vidal, 2003), which find important applications to quantum information systems.

Chapter 7

Optimal Interpolation Norms

In Chapter 2 we reviewed a range of structured sparsity penalties, and we highlighted that norms defined by a variational formulation have increasingly been studied, enjoying strong estimation properties over numerous learning problems. The vector k-support norm is a particular example of such penalties, deriving from the group lasso with overlap of Jacob et al. (2009b). In the previous chapters we generalized the norm to apply to matrices (Chapter 4) and tensors (Chapter 6), and numerical experiments showed statistically significant improvement over the standard regularizers in the literature. An extension of the k-support norm to introduce a parameter p further allows the penalty to be tuned, which in the matrix case corresponds to tuning the decay of the singular values.

Each of the k-support norms is perhaps most easily characterized via its unit ball. In the case of the vector (k,p)-support norm, its unit ball corresponds to the convex hull of vectors with cardinality k and unit ℓ_p-norm, with similar constructions in the matrix and tensor frameworks. However, as we have seen, the norms also admit a characterization as an infimal convolution. For example, for the (k,p)-support norm of Chapter 5 (Definition 5.4) we have

$$\|w\|_{(k,p)} = \inf_{(v_g)} \left\{ \sum_{g \in \mathcal{G}_k} \|v_g\|_p \;\middle|\; \sum_{g \in \mathcal{G}_k} v_g = w \right\}. \tag{7.1}$$

where the infimum is over all vectors $v_g \in \mathbb{R}^d$ such that $\operatorname{supp}(v_g) \subseteq g$, for $g \in \mathcal{G}_k$.

In this chapter we define a general framework for norms which generalizes the infimal convolution construction in (7.1). Specifically, we propose a class of norms defined via an optimal interpolation problem involving the composition of norms and a linear operator. This framework is very general and encompasses a number of norms which have been used as regularizers in machine learning and statistics, including the k-support norms and allows us to construct more via an optimal interpolation problem involving simpler norms and a linear operator.

Specifically, the norm of $w \in \mathbb{R}^d$ is defined as

$$\|w\| = \inf_{\substack{v\in\mathbb{R}^N \\ Bv=w}} |||F(v)|||, \tag{7.2}$$

where $|||\cdot|||$ is a monotone norm (e.g., an ℓ_p norm), F a vector-valued mapping the components of which are also norms, and B a $d \times N$ real matrix. As we shall see, this concise formulation encompasses many classes of regularizers, either via the norm (7.2) or the corresponding dual norm, which we investigate in this chapter. Furthermore, while we focus on norms, the construction (7.2) provides a scheme for a broader range of regularizers of the form

$$\varphi(w) = \inf_{\substack{v\in\mathbb{R}^N \\ Bv=w}} h\big(F(v)\big), \tag{7.3}$$

where the components of F and h are convex functions that satisfy certain properties. Such functions generalize the notion of infimal convolution (see Bauschke and Combettes, 2011, Proposition 12.35) and are found in signal recovery formulations, see, for example, Becker and Combettes (2014).

Optimal interpolation and, in particular, the problem of finding a minimal norm interpolant to a finite set of points has a long history in approximation theory; see, for example, Cheney and Light (2000) and the references therein. A special case of this setup occurs when we do minimal norm interpolation in a Hilbert space. Specifically, we choose the Sobolev space

$$H^s(\mathbb{R}^d) = \left\{ f \in L^2(\mathbb{R}^d) \;\middle|\; \|f\|_{H^s} = \sqrt{(2\pi)^{-n}\int_{\mathbb{R}^d}(1+|\xi|^2)^s|\hat{f}(\xi)|^2 d\xi} < \infty \right\},$$

where $\hat{f}$ denotes the Fourier transform of $f \in L^2(\mathbb{R}^d)$. The norm $\|\cdot\|_{H^s}$ makes $H^s(\mathbb{R}^d)$ into a reproducing kernel Hilbert space, and so, the solution of the minimal interpolation problem with function evaluation constraints on a finite set of points $x_1,\ldots,x_m \in \mathbb{R}^d$, yields an interpolant which is formed from the reproducing kernel of $H^s(\mathbb{R}^d)$.

In machine learning, the problem (7.3) has been investigated in Argyriou et al. (2010) and Micchelli and Pontil (2005). Special cases of our setting based on (7.2) appear in Maurer and Pontil (2012); Tomioka and Suzuki (2013), where $|||\cdot|||$ is an ℓ_p norm and the components of F are ℓ_2 norms. On the other hand, the atomic norms considered in Chandrasekaran et al. (2012b) correspond, in the case of a finite number of atoms, to setting $F = \mathrm{Id}$ and $|||\cdot||| = \|\cdot\|_1$ in (7.2). The general fact that the construction (7.2) preserves convexity is known. However, to the best of our knowledge, the observation that optimal interpolation can be used as a framework to generate norms is novel. On the numerical front, we shall take advantage of the fact that the mapping F and the operator B can be composed of a large number of “simple” components to create complex structures, leading to efficient algorithms to solve problems involving regularizers of the form (7.2).

The remainder of this chapter is organized as follows. In Section 7.1 we introduce the general class of regularizers and establish some of their basic properties. In Section 7.2 we give a number of examples of norms in this framework. In Section 7.3 we provide a random block-coordinate algorithm to solve learning problems involving these penalties. In Section 7.4 we report on numerical experiments with this algorithm.

7.1 A Class of Norms

We first establish the mathematical foundation of our framework. We begin with a scheme to construct convex functions on $\mathbb{R}^d$.

Proposition 84. *Let m, d, and N be strictly positive integers. Let $K = \mathbb{R}^m_-$, let $B \in \mathbb{R}^{d\times N}$, and let $F\colon \mathbb{R}^N \to \mathbb{R}^m$ be K-convex in the sense that, for every $\alpha \in \left]0,1\right[$ and every $(u,v) \in \mathbb{R}^N \times \mathbb{R}^N$*

$$F\big(\alpha u + (1-\alpha)v\big) - \alpha F(u) - (1-\alpha)F(v) \in K. \tag{7.4}$$

Let $h\colon \mathbb{R}^m \to \left]-\infty,+\infty\right]$ be a proper convex function such that, for every $(x,y) \in \operatorname{ran} F \times \operatorname{ran} F$,

$$x - y \in K \quad \Rightarrow \quad h(x) \le h(y). \tag{7.5}$$

Define, for every $w \in \mathbb{R}^d$,

$$\varphi(w) = \inf_{\substack{v\in\mathbb{R}^N\\ Bv=w}} h\big(F(v)\big). \tag{7.6}$$

Then $h \circ F$ and φ are convex.

Proof. It is enough to show that $h\circ F$ is convex, as this will imply that φ is likewise by Bauschke and Combettes (2011, Proposition 12.34(ii)) or Rockafellar (1970, Theorem 5.7). Let $\alpha \in \left]0,1\right[$ and let u and v be points in $\mathbb{R}^N$. Combining (7.4) and (7.5) yields

$$h\big(F(\alpha u + (1-\alpha)v)\big) \le h(\alpha F(u) + (1-\alpha)F(v)). \tag{7.7}$$

Therefore, by convexity of h, we obtain

$$(h\circ F)\big(\alpha u + (1-\alpha)v\big) \le \alpha(h\circ F)(u) + (1-\alpha)(h\circ F)(v), \tag{7.8}$$

which establishes the convexity of $h\circ F$. □

We are now ready to define the class of norms induced by optimal interpolation. As we shall see below, the construction involves several norms and their duals in a manner which produces a very flexible framework. In order to minimize confusion due to notation, throughout this chapter we refer to the principle norm as $|||\cdot|||$, and the constituent norms will be given by f and g, with duals denoted respectively by $|||\cdot|||_*$, f_* and g_*.

Assumption 85. *Let m, d, and $(r_j)_{1\leq j\leq m}$ be strictly positive integers, set $N=\sum_{j=1}^m r_j$, and let $v=(v_j)_{1\leq j\leq m}$ denote a generic element in $\mathbb{R}^N$ where, for every $j\in\{1,\ldots,m\}$, $v_j\in\mathbb{R}^{r_j}$. Furthermore:*

1. *$F\colon v\mapsto (f_j(v_j))_{1\leq j\leq m}$ where, for every $j\in\{1,\ldots,m\}$, f_j is a norm on $\mathbb{R}^{r_j}$ with dual norm $f_{j,*}$.*

2. *$|||\cdot|||$ is a norm on $\mathbb{R}^m$ which is monotone in the sense that, for every $(x,y)\in\mathbb{R}^m_+\times\mathbb{R}^m_+$,*

$$x-y\in\mathbb{R}^m_- \quad\Rightarrow\quad |||x|||\leq |||y|||, \tag{7.9}$$

and $|||\cdot|||_*$ denotes its dual norm.

3. *For every $j\in\{1,\ldots,m\}$, $B_j\in\mathbb{R}^{d\times r_j}$, and $B=[B_1\ \cdots\ B_m]$ has full rank.*

Let, for every $w\in\mathbb{R}^d$,

$$\|w\|=\min_{\substack{v\in\mathbb{R}^N\\ Bv=w}}|||F(v)|||=\min_{\substack{v\in\mathbb{R}^N\\ \sum_{j=1}^m B_jv_j=w}}\big|\big|\big|\big(f_1(v_1),\ldots,f_m(v_m)\big)\big|\big|\big|. \tag{7.10}$$

Proposition 86. *Consider the setting of Assumption 85. Then $|||\cdot|||\circ F$ is a norm on $\mathbb{R}^N$ and its dual norm at $t\in\mathbb{R}^N$ is*

$$(|||\cdot|||\circ F)_*(t)=\big|\big|\big|\big(f_{1,*}(t_1),\ldots,f_{m,*}(t_m)\big)\big|\big|\big|. \tag{7.11}$$

Proof. We show that $|||\cdot|||\circ F$ satisfies the properties of a norm and compute its dual norm, see Appendix F. □

Remark 87. *Let $w\in\mathbb{R}^d$, set $C=\big\{v\in\mathbb{R}^N\ \big|\ Bv=w\big\}$, and let d_C be the distance function to C associated with the norm $|||\cdot|||\circ F$ (see Proposition 86), that is,*

$$(\forall z\in\mathbb{R}^N)\quad d_C(z)=\inf_{v\in C}|||F(v-z)|||. \tag{7.12}$$

Note that C is a closed affine subspace and it follows from (7.12) *and* (7.10) *that*

$$d_C(0)=\inf_{v\in C}|||F(v-0)|||=\inf_{v\in C}|||F(v)|||=\|w\|. \tag{7.13}$$

Thus, the function $\|\cdot\|$ in (7.10) is defined via a minimal norm interpolation process. Hence, the optimization problem underlying (7.10) is that of minimizing the norm $|||\cdot|||\circ F$ over the closed affine subspace C. It therefore possesses a solution, namely the minimum norm element in C.

In the next result we establish that the construction described in Assumption 85 does provide a norm, and we compute its dual norm.

Theorem 88. *Consider the setting of Assumption 85. Then the following hold:*

1. *$\|\cdot\|$ is a norm.*
2. *The dual norm of $\|\cdot\|$ at $u\in\mathbb{R}^d$ is*

$$\|u\|_* = \Big|\Big|\Big|\big(f_{1,*}(B_1^\top u),\dots,f_{m,*}(B_m^\top u)\big)\Big|\Big|\Big|_*. \tag{7.14}$$

Proof. We show that the function $\|\cdot\|$ satisfies the properties of the norm and compute its dual norm, see Appendix F. □

A special case of Theorem 88 is the following corollary, which appears in Maurer and Pontil (2012, Theorem 7).

Corollary 89. *Let p, q, r, and s be numbers in $[1,+\infty]$ such that $1/p+1/q=1$ and $1/r+1/s=1$. Then the function $\|\cdot\|$ defined at $w\in\mathbb{R}^d$ by*

$$\|w\| = \min_{\substack{v\in\mathbb{R}^N\\ \sum_{j=1}^m B_j v_j = w}} \|(\|v_1\|_p,\dots,\|v_m\|_p)\|_r \tag{7.15}$$

is a norm and the dual norm at $u\in\mathbb{R}^d$ is given by

$$\|u\|_* = \|(\|B_1^\top u\|_q,\dots,\|B_m^\top u\|_q)\|_s. \tag{7.16}$$

Proof. This construction is a special case of Assumption 85 with $|||\cdot|||=\|\cdot\|_r$ and $f_1=\dots=f_m=\|\cdot\|_p$. The function $|||\cdot|||$ is an absolute norm, that is, $|||\cdot|||=|||\cdot|||\circ\Lambda$, for every $\Lambda=\mathrm{diag}(\lambda_1,\dots,\lambda_m)$ such that $(\forall j\in\{1,\dots,m\})\ \lambda_j\in\{-1,1\}$. By Horn and Johnson (1991, Theorem 5.4.19) a norm is monotone if and only if it is absolute. The result follows by applying Theorem 88. □

Remark 90. *Suppose that, in Proposition 84, F is the identity mapping and h an arbitrary norm. Then, arguing as in the proof of Theorem 88, we obtain that the function φ of (7.6) is a norm and that the dual norm at $u\in\mathbb{R}^d$ is $|||B^\top u|||_*$.*

Remark 91. *Any norm $\|\cdot\|$ can trivially be written in the form of (7.6), by simply letting $|||\cdot|||=\|\cdot\|$, F be the identity operator, and B be the identity matrix. However we are interested* *in exploiting the structure of the construction* (7.6) *in cases in which the norms* $|||\cdot|||$ *and* $(f_j)_{1\leq j\leq m}$ *are chosen from a "simple" class and give rise, via the optimal interpolation problem* (7.6)*, to a "complex" norm* $\|\cdot\|$ *which we shall use as a regularizer in a statistical learning setting. In particular, when using proximal splitting methods, the computation of* $\mathrm{prox}_{\|\cdot\|}$ *will typically not be easy whereas that of* $\mathrm{prox}_{|||\cdot|||}$ *and* prox_F *will be. This will be exploited in Section 7.3 to devise efficient splitting algorithms to solve problems involving* $\|\cdot\|$ *which use the functions* $(f_j)_{1\leq j\leq m}$ *and the operators* $(B_j)_{1\leq j\leq m}$ *separately.*

7.2 Examples

In this section, we observe that the construct presented in Section 7.1 encompasses in a unified framework a number of existing regularizers. For simplicity, we focus on the norms captured by Corollary 89. Our main aim here is not to derive new regularizers but, rather, to show that our analysis captures existing ones and to derive their dual norms.

A number of examples that we present in this section were discussed in Chapter 2. In order to show how these norms fit into the framework introduced in this chapter, we recall the definitions and we adapt the notation to the framework presented here.

Given $w\in\mathbb{R}^d$ and $J\subset\{1,\ldots,d\}$, recall that $w_{|J}\in\mathbb{R}^d$ is the vector w restricted to the support set defined by J, where the support of w is $\mathrm{supp}(w)=\big\{i\;\big|\;1\leq i\leq d\;,w_i\neq 0\big\}$.

7.2.1 Latent group lasso

The first example we consider is known as the latent group lasso (LGL), or group lasso with overlap, which goes back to Jacob et al. (2009b). Let $(G_j)_{1\leq j\leq m}$ be nonempty subsets of $\{1,\ldots,d\}$ such that $\bigcup_{j=1}^m G_j=\{1,\ldots,d\}$ and define the vector space

$$Z=\big\{z=(z_j)_{1\leq j\leq m}\;\big|\;(\forall j\in\{1,\ldots,m\})\;z_j\in\mathbb{R}^d\text{ and }\mathrm{supp}(z_j)\subset G_j\big\}. \tag{7.17}$$

The latent group lasso penalty is defined, for $w\in\mathbb{R}^d$, as

$$\|w\|_{\mathrm{LGL}}=\min\left\{\sum_{j=1}^m\|z_j\|_p\;\middle|\;z\in Z,\;\sum_{j=1}^m z_j=w\right\}. \tag{7.18}$$

The optimal interpolation problem (7.18) seeks a decomposition of vector w in terms of vectors $(z_j)_{1\leq j\leq m}$ whose support sets are restricted to the corresponding group of variables in G_j. If the groups overlap then the decomposition is not necessarily unique, and the variational formulation involves those z_j such that the aggregate ℓ_p norm is minimal. On the other hand, if the group are pairwise disjoint, that is $\{G_1,\ldots,G_m\}$ forms a partition of $\{1,\ldots,d\}$, the latent group lasso coincides with the "standard" group lasso (Yuan and Lin, 2006), which is defined, for $w \in \mathbb{R}^d$, as

$$\|w\|_{\mathrm{GL}} = \sum_{j=1}^{m} \|w_{|G_j}\|_p. \tag{7.19}$$

In the general case (7.18) has no closed form expression due to the overlapping of the groups. However, in special cases which exhibit additional structure, it can be computed in a finite number of steps; an important example is provided by the (k,p)-support norm (McDonald et al., 2016b), in which the groups consists of all subsets of $\{1,\dots,d\}$ of cardinality no greater than k, for some $k \in \{1,\dots,d\}$, see Chapter 5. The case $p = 2$ has been studied in Argyriou et al. (2012) and McDonald et al. (2014a), see Chapter 3.

Proposition 92. *For every $j \in \{1,\dots,m\}$, set $p_j = |G_j|$. The latent group lasso penalty* (7.18) *is a norm of the form* (7.10) *with $f_j = \|\cdot\|_p$, $|||\cdot||| = \|\cdot\|_1$, and $d \times p_j$ matrices $B_j = [e_i \mid i \in G_j]$, where e_i is the ith standard unit vector in $\mathbb{R}^d$. Furthermore, the dual norm of $w \in \mathbb{R}^d$ is given by*

$$\|w\|_{\mathrm{LGL},*} = \max_{1 \le j \le m} \|w_{|G_j}\|_q.$$

Proof. The expression for the primal norm follows immediately by the change of variable $z_j = B_j v_j$, which also eliminates the support constraint. The dual norm follows by (7.14) noting that $|||\cdot|||_*$ is the max norm, $f_{j,*} = \|\cdot\|_q$ for all j and $B_j^\top w = B_j w = w_{|G_j}$. □

7.2.2 Overlapping group lasso

An alternative generalization of the group lasso is the overlapping group lasso (Jenatton et al., 2011b; Zhao et al., 2009), which we write as $\|\cdot\|_{\mathrm{OGL}}$. It has the same expression of the group lasso in (7.19), except that we drop the restriction that the set of groups $\{G_1,\dots,G_m\}$ form a partition of the index set $\{1,\dots,d\}$. That is, we require only that the union of sets G_j be equal to $\{1,\dots,d\}$. Our next result establishes that the overlapping group lasso penalty is captured by a dual norm of type (7.14).

Proposition 93. *Let $p \in [1,+\infty]$. Let $r_j = |G_j|$, $j \in \{1,\dots,m\}$, let $f_j = \|\cdot\|_q$, let $|||\cdot||| = \|\cdot\|_\infty$, and the $d \times r_j$ matrices $B_j = [e_i \mid i \in G_j]$. Then then norm* (7.10) *evaluated at $w \in \mathbb{R}^d$ is*

$$\|w\| = \inf\left\{ \max_{1 \le j \le m} \|v_j\|_q \;\middle|\; \sum_{j=1}^{m} B_j v_j = w \right\}$$

and $\|\cdot\|_* = \|\cdot\|_{\mathrm{OGL}}$.

Proof. Note that $|||\cdot|||_* = \|\cdot\|_1$ and $(\forall j \in \{1,\dots,m\})$ $f_{j,*} = \|\cdot\|_p$. Therefore, the duality assertion follows from (7.14) by noting that $\|B_j^\top u\|_p = \|u_{|G_j}\|_p$. □

The case $p = +\infty$ corresponds to the iCAP penalty in Zhao et al. (2009). We may also consider other choices for the matrices B_j. For example, an appropriate choice gives various total variation penalties Micchelli et al. (2011). A further example is obtained by choosing $m = 1$ and B_1 to be the incidence matrix of a graph, a setting which is relevant in online learning over graphs Herbster and Lever (2009). In particular, for $p = 1$, this corresponds to the fused lasso penalty Tibshirani and Saunders (2005).

7.2.3 The ordered weighted ℓ_1-norm

The ordered weighted ℓ_1-norm (OWL) was introduced for regression problems with correlated coefficients (Bogdan et al., 2013; Zeng and Figueiredo, 2014; Figueiredo and Nowak, 2016). It is defined, for $w \in \mathbb{R}^d$, as

$$\|w\|_{\text{OWL}} = \left\langle \lambda \mid |w|^{\downarrow} \right\rangle, \tag{7.20}$$

where $\lambda \in \mathbb{R}^d_+$ is a vector of weights such that $\lambda_1 \geq \ldots \geq \lambda_d$, and $\lambda_1 > 0$. The norm satisfies $\lambda_1 \|w\|_\infty \leq \|w\|_{\text{OWL}} \leq \lambda_1 \|w\|_1$, with equalities when $\lambda_2 = \ldots = \lambda_d = 0$, or $\lambda_1 = \ldots = \lambda_d$, respectively. Our next result establishes that the ordered weighted ℓ_1-norm is captured by a dual norm of type (7.14).

Proposition 94. *Let $\lambda \in \mathbb{R}^d_+$ such that $\lambda_1 \geq \ldots \geq \lambda_d$ and these are not all zero, and let $n = d!$. The ordered weighted ℓ_1-norm defined by (7.20) is a norm of the form (7.14) with $f_j = \|\cdot\|_\infty$, $|||\cdot||| = \|\cdot\|_1$, and matrices B_j be the $d \times d$ matrices given by $B_j = diag(\lambda^j)$, where λ^j is the j-th permutation of $(\lambda_1, \ldots, \lambda_d)$.*

Proof. Let Π be the set of permutations on $\mathbb{R}^d$ and let $\pi^* \in \Pi$ be the permutation defined by $\pi^* : |u|^{\downarrow} \mapsto |u|$. We have

$$\begin{aligned}\|u\|_{\text{OWL}} = \left\langle \lambda \mid |u|^{\downarrow} \right\rangle &= \max_{\pi \in \Pi} \left\langle \pi(\lambda) \mid |u|^{\downarrow} \right\rangle = \max_{\pi \in \Pi} \left\langle \pi^*(\pi(\lambda)) \mid \pi^*(|u|^{\downarrow}) \right\rangle \\ &= \max_{\pi' \in \Pi} \langle \pi'(\lambda) \mid |u| \rangle = \max_{j=1}^{n} \langle \lambda^j \mid |u| \rangle = \max_{j=1}^{n} \|B^j u\|_1,\end{aligned}$$

as required, where we have used the Hardy-Littlewood-Polya inequality, Theorem 10, which states that any vectors $x, y \in \mathbb{R}^d$ satisfy the inequality $\langle x \mid y \rangle \leq \left\langle x^{\downarrow} \mid y^{\downarrow} \right\rangle$, the fact that $\langle x \mid y \rangle = \langle \pi(x) \mid \pi(y) \rangle$ for any $\pi \in \Pi$ and the fact that Π is closed under composition. □

7.2.4 Θ-norms

In Chapter 3 we considered the Θ-norms, a family of norms which is parameterized by a convex set; see also Bach et al. (2012); Micchelli et al. (2013). We consider the generalized form introduced in Section 3.4. Specifically, let $\Theta \subset \mathbb{R}^d_{++}$ be nonempty, convex, and bounded, let $p \in [1, \infty]$ and define $q \in [1, \infty]$ by $\frac{1}{p} + \frac{1}{q} = 1$, and define $\phi_p : \mathbb{R}_+ \mapsto \mathbb{R}_+$ by $\phi_p(x) = x^p$.

Then the expressions

$$\inf_{\theta \in \Theta} \phi_p^{-1}\left(\sum_{i=1}^{d} \theta_i \phi_p\left(\frac{|w_i|}{\theta_i}\right)\right) \tag{7.21}$$

and

$$\sup_{\theta \in \Theta} \phi_q^{-1}\left(\sum_{i=1}^{d} \theta_i \phi_q(|u_i|)\right) \tag{7.22}$$

define norms, which we refer to as the primal and dual norms, respectively. Norms which are included in this family include the ℓ_p-norms, and the (k,p)-support norm. We now consider the case when Θ is a polyhedron.

Proposition 95. *Let Θ be a nonempty subset of $\mathbb{R}^d_{++}$ such that $\bar{\Theta} = \operatorname{conv}\{\theta^{(1)},\ldots,\theta^{(m)}\}$, where $(\forall j \in \{1,\ldots,m\})\ \theta^{(j)} \in \mathbb{R}^d_{+}$. Then the norms defined in* (7.21) *and* (7.22) *can be written in the form* (7.10) *and* (7.14) *respectively, with $|||\cdot||| = \|\cdot\|_1$ and $(\forall j \in \{1,\ldots,m\})\ f_j = \|\cdot\|_p$ and $B_j = \operatorname{diag}\left((\theta^{(j)})^{\frac{1}{q}}\right)$.*

Proof. For every $j \in \{1,\ldots,m\}$, set $B_j = \operatorname{diag}\left((\theta^{(j)})^{\frac{1}{q}}\right)$ and let $f_{j,*}$ be the ℓ_q-norm on $\mathbb{R}^d$. Furthermore, let $|||\cdot|||_*$ be the ℓ_∞-norm on $\mathbb{R}^m$. Then the dual norm of $\|\cdot\|$ at $u \in \mathbb{R}^d$ is derived from (7.14) to be

$$\begin{aligned}
\|u\|_* &= \max_{1 \le j \le m} \|B_j^\top u\|_p \\
&= \max_{1 \le j \le m} \left(\sum_{i=1}^{d} \theta_i^{(j)} |u_i|^q\right)^{\frac{1}{q}} \\
&= \max_{\theta \in \bar{\Theta}} \left(\sum_{i=1}^{d} \theta_i |u_i|^q\right)^{\frac{1}{q}} \qquad (7.23)\\
&= \sup_{\theta \in \Theta} \left(\sum_{i=1}^{d} \theta_i |u_i|^q\right)^{\frac{1}{q}},
\end{aligned}$$

where equality (7.23) uses the fact that a linear function on a compact convex set attains its maximum at an extreme point of the set. It follows that the dual norms coincide, and consequently the same holds for the primal norms. □

7.2.5 Polyhedral norms

Polyhedral norms (Hiriart-Urruty and Lemaréchal, 1993) are characterized by having as their unit ball a polyhedron, that is, a finite intersection of closed affine halfspaces. Specifically, let m be an even strictly positive integer and let $E = \{b_1,\ldots,b_m\} \subset \mathbb{R}^d$ be such that span$(E) = \mathbb{R}^d$ and $E = -E$. Let $C = \text{conv}(E)$ and let $\|\cdot\|_{\text{PH}}$ be the polyhedral norm with unit ball C. We have the following characterization of a polyhedral norm (Körner, 2011, Section 1.34),

$$\|w\|_{\mathrm{PH}} = \min\left\{ \sum_{j=1}^{m} |v_j| \;\Big|\; v \in \mathbb{R}^m,\; \sum_{j=1}^{m} v_j b_j = w \right\}. \tag{7.24}$$

Using this expression we have the following result, which follows directly by Remark 90.

Proposition 96. *The polyhedral norm on $\mathbb{R}^d$ defined by* (7.24) *can be written in the form of* (7.10) *for $r_1 = \cdots = r_m = 1$, $B_j = [b_j]$, $|||\cdot||| = \|\cdot\|_1$, and F the identity mapping.*

Recall that the polar of a set C is the set $C^{\odot} = \{u \in \mathbb{R}^d \mid \sup_{w \in C} \langle u \mid w \rangle \leq 1\}$. The dual norm $\|\cdot\|_{\mathrm{PH},*}$ is also polyhedral, and its unit ball is the convex hull of $E^{\ominus} = \{b_1^{\ominus}, \ldots, b_r^{\ominus}\}$, where $E^{\ominus}$ is the polar cone of E and r is an even positive integer. Indeed, the dual norm at $u \in \mathbb{R}^d$ is given by

$$\|u\|_{\mathrm{PH},*} = \max\{\langle u \mid w \rangle \mid w \in C\} = \max_{1 \leq j \leq m} \langle u \mid b_j \rangle. \tag{7.25}$$

It follows that $\|w\|_{\mathrm{PH}}$ can also be written as a maximum of scalar products over the set of extreme points of the unit ball of the dual norm, that is, $\|w\|_{\mathrm{PH}} = \max\{\langle w \mid b \rangle \mid b \in E^{\ominus}\}$, where $E^{\ominus}$ is the set of extreme points of the dual unit ball $C^{\odot}$. Geometrically, the unit balls are related by polarity.

7.2.6 Tensor norms

Recently a number of regularizers have been proposed to learn low rank tensors. We discuss three of them. The first two are based on the nuclear norm of the matricizations of a tensor.

Recall that the nuclear norm (or trace norm) of a matrix, $\|\cdot\|_{\mathrm{nuc}}$, is the sum of its singular values. Its dual norm $\|\cdot\|_{\mathrm{sp}}$ is given by the largest singular value. The *overlapped nuclear norm* (Romera-Paredes and Pontil, 2013; Tomioka et al., 2011) is defined as the sum of the nuclear norms of the mode-j matricizations, namely

$$\|\boldsymbol{W}\|_{\mathrm{ONN}} = \sum_{j=1}^{m} \|M_j(\boldsymbol{W})\|_{\mathrm{nuc}}. \tag{7.26}$$

The next result, which is easily proved, captures the overlapped nuclear norm in our framework.

Proposition 97. *The overlapped nuclear norm* (7.26) *can be written in the form* (7.14) *for $|||\cdot|||_* = \|\cdot\|_1$, $f_{j,*} = \|\cdot\|_{\mathrm{nuc}}$ and $B_j^\top = M_j$. Furthermore, the dual norm of a matrix $\boldsymbol{U}$ is*

$$\|\boldsymbol{U}\|_{\mathrm{ONN},*} = \inf\left\{ \max_{1 \leq j \leq m} \|V_j\|_{\mathrm{sp}} \;\Big|\; \sum_{j=1}^{m} M_j^\top V_j = \boldsymbol{U} \right\}. \tag{7.27}$$

Let $\alpha_1,\ldots,\alpha_m$ be strictly positive reals. The *scaled latent nuclear norm* is defined by variational problem

$$\|\boldsymbol{W}\|_{\text{LNN}} = \inf\left\{\sum_{j=1}^{m}\frac{1}{\alpha_j}\|V_j\|_{\text{nuc}} \,\Big|\, \sum_{j=1}^{m} M_j^{\top} V_j = \boldsymbol{W}\right\}. \tag{7.28}$$

The case $\alpha_j = 1$ is presented in Tomioka and Suzuki (2013) and the case $\alpha_j = \sqrt{d_j}$ in Wimalawarne et al. (2014).

Proposition 98. *The latent nuclear norm* (7.28) *can be written in the form* (7.10) *for* $|||\cdot||| = \|\cdot\|_1$*,* $f_j = \|\cdot\|_{\text{nuc}}/\alpha_j$ *and* $B_j = M_j^{\top}$*. Furthermore, the dual norm of a matrix* $\boldsymbol{U}$ *is*

$$\|\boldsymbol{U}\|_{\text{LNN},*} = \max_{1\le j\le m} \alpha_j \|M_j(\boldsymbol{U})\|_{\text{sp}}.$$

Proof. Immediate, upon identification of $|||\cdot|||$, F, the matrices $(B_j)_{1\le j\le m}$, and noting that $(\forall j \in \{1,\ldots,m\})\ f_{j,*} = \alpha_j\|\cdot\|_{\text{sp}}$. □

The final tensor norm that we consider is the tensor k-support norm that we defined in Chapter 6. Recall from Equation (6.7) that the primal and dual norms are defined, for $\boldsymbol{W},\boldsymbol{U} \in \mathbb{R}^{d_1\times\cdots\times d_m}$, respectively, as

$$|||\boldsymbol{W}||| = \inf\left\{\sum_{j=1}^{m}\|M_j(\boldsymbol{V}_j)\|_{(k_j,p_j)} \,\Big|\, \sum_{j=1}^{m}\boldsymbol{V}_j = \boldsymbol{W}\right\} \tag{7.29}$$

$$|||\boldsymbol{U}|||_* = \max_{1\le j\le m}\|M_j(\boldsymbol{U})\|_{(k_j,p_j),*} \tag{7.30}$$

where the infimum in the primal equation is over the tensors $\boldsymbol{V}_j \in \mathbb{R}^{d_1\times\cdots\times d_m}$ for $j \in \{1,\ldots,m\}$.

Proposition 99. *The tensor k-support norm can be written in the form* (7.10) *for* $|||\cdot||| = \|\cdot\|_1$*,* $f_j = \|\cdot\|_{(k_j,p_j)}$ *and* $B_j = M_j^{\top}$*.*

Proof. Immediate from the dual norm. □

7.2.7 Norms induced by reproducing kernels

We give an infinite-dimensional example. Let X be a set; for every every $j \in \{1,\ldots,m\}$, let $K_j : X\times X \to \mathbb{R}$ be a positive semidefinite kernel, denote by H_j the reproducing kernel Hilbert space (RKHS) associated to K_j,and denote by $\|\cdot\|_j$ the norm induced by the inner product in H_j. For background on RKHS's we refer to Aronszajn (1950).

Let $H = \cup_{1\le j\le m} H_j$. Using our framework we can endow H with the norm, defined at $h \in H$, as

$$\|h\| = \left\{|||(\|h_1\|_1,\ldots,\|h_m\|_m)||| \,\Big|\, \sum_{j=1}^{m} h_j = h\right\}. \tag{7.31}$$

The very special case that $m = 2$ and $|||\cdot|||$ is the ℓ_2 norm is discussed in (Aronszajn, 1950, p. 353). The important case that $|||\cdot|||$ is the ℓ_1 norm has been considered in a number of papers, in which (7.31) is employed as a regularizer in multiple kernel learning problems, see for example Bach et al. (2012) and references therein. The extension where $|||\cdot|||$ is the ℓ_p norm is presented in Micchelli and Pontil (2005); Kloft et al. (2011).

It is interesting to note that, for a special choice of the kernels $K_1, \ldots, K_m$, the norm (7.31) includes the latent group lasso norm described earlier, see also Bach et al. (2012); Micchelli et al. (2013) for related observations. Specifically, we choose $X = \mathbb{R}^d$, and $(\forall j \in \{1, \ldots, m\}$, $\forall (x,z) \in \mathbb{R}^d \times \mathbb{R}^d)$ let $K_j(x,z) = \langle x \mid P_j z \rangle$, where P_j is the projection to the linear span of the standard coordinates vectors associated to the index set $G_j \subset \{1, \ldots, d\}$. For this choice, H_j is the space of linear functions $x \mapsto \langle w_j \mid P_j x \rangle$, for some $w_j \in \mathbb{R}^d$. Therefore setting $v_j = P_j w_j$, we see that (7.31) rewrites in the form of (7.18) for $p = 2$.

7.3 Random Block-Coordinate Algorithm

The purpose of this section is to address some of the numerical aspects associated with the class of norms introduced in Assumption 85. Since such norms are nonsmooth convex functions, they could in principle be handled via their proximity operators in the context of proximal splitting algorithms (Bauschke and Combettes, 2011). However, the proximity operator of the composite norm $\|\cdot\|$ in (7.10) is usually too complicated for this direct approach to be viable. We circumvent this problem by formulating the problem in such a way that it involves only the proximity operators of the functions $(f_j)_{1 \leq j \leq m}$, which will typically be available in closed form. The main features and novelty of the algorithmic approach we propose are the following:

- It can handle general nonsmooth formulations: the functions present in the model need not be differentiable.

- It adapts the recent approach proposed in Combettes and Pesquet (2015) to devise a block-coordinate algorithm which allows us to select arbitrarily the blocks functions $(f_j)_{1 \leq j \leq m}$ to be activated over the course of the iterations. This makes the method amenable to the processing of very large data sets in a flexible manner by adapting the computational load of each iteration to the available computing resources.

- The computations are broken down to the evaluation of simple proximity operators of the functions f_j and of those appearing in the loss function, while the linear operators are applied separately.

- It guarantees the convergence of the iterates to a solution of the minimization problem under consideration; convergence of the function values is also guaranteed.

We are not aware of alternative frameworks which achieve simultaneously these properties. In particular, it is noteworthy that the convergence of the iterates is achieved although only a portion of the proximity operators are activated at each iteration.

We consider a supervised learning problem in which a vector $w \in \mathbb{R}^d$ is to be inferred from n input-output data points $(a_i, \beta_i)_{1 \leq i \leq n}$, where $a_i \in \mathbb{R}^d$ and $\beta_i \in \mathbb{R}$. A common formulation for such problems is the regularized convex minimization problem

$$\underset{w \in \mathbb{R}^d}{\text{minimize}} \; \sum_{i=1}^{n} \ell_i(\langle w \mid a_i \rangle, \beta_i) + \lambda \|w\|, \tag{7.32}$$

where $\ell_i \colon \mathbb{R}^2 \to \mathbb{R}$ is a loss function (for simplicity we take it to be real-valued but extended functions can also be considered) such that, for every $\beta \in \mathbb{R}$, $\ell_i(\cdot, \beta)$ is convex, $\lambda \in \left]0, +\infty\right[$ is a regularization parameter, and $\|\cdot\|$ assumes the general form (7.10). For simplicity, we focus in this section and the subsequent experiments on the case when $|||\cdot||| = \|\cdot\|_1$. As discussed in Section 7.2, this choice covers many instances of practical interest in machine learning. We can rewrite (7.32) as

$$\underset{\substack{w \in \mathbb{R}^d \\ v_1 \in \mathbb{R}^{r_1}, \ldots, v_m \in \mathbb{R}^{r_m} \\ \sum_{j=1}^m B_j v_j = w}}{\text{minimize}} \; \sum_{i=1}^{n} \ell_i(\langle w \mid a_i \rangle, \beta_i) + \lambda \sum_{j=1}^{m} f_j(v_j). \tag{7.33}$$

We can therefore solve the optimization problem

$$\underset{v_1 \in \mathbb{R}^{r_1}, \ldots, v_m \in \mathbb{R}^{r_m}}{\text{minimize}} \; \sum_{i=1}^{n} \ell_i\Bigg(\sum_{j=1}^{m} \langle B_j v_j \mid a_i \rangle, \beta_i \Bigg) + \lambda \sum_{j=1}^{m} f_j(v_j), \tag{7.34}$$

and derive from one of its solutions $(v_j)_{1 \leq j \leq m}$ a solution $w = \sum_{j=1}^m B_j v_j$ to (7.33). The algorithm adapts the block-coordinate approach of Combettes and Pesquet (2015) to our setting, with a novel update step for the loss function. We present the full derivation in Appendix F, and outline the key steps below. We first introduce the functions

$$\Phi \colon (v_1, \ldots, v_m) \mapsto \lambda \sum_{j=1}^{m} f_j(v_j) \tag{7.35}$$

and

$$\Psi \colon (\eta_1, \ldots, \eta_n) \mapsto \sum_{i=1}^{n} \psi_i(\eta_i), \tag{7.36}$$

where, for every $i \in \{1, \ldots, n\}$,

$$\psi_i \colon \eta_i \mapsto \ell_i(\eta_i, \beta_i). \tag{7.37}$$

We next designate by $A \in \mathbb{R}^{n\times d}$ the matrix the rows of which are $(a_i)_{1\leq i\leq n}$ and define

$$L = [L_1 \cdots L_m] \in \mathbb{R}^{n\times N}, \tag{7.38}$$

where, for every $j \in \{1,\ldots,m\}$,

$$L_j = AB_j \in \mathbb{R}^{n\times r_j}. \tag{7.39}$$

Then $L = AB$, where $B = [B_1 \cdots B_m]$, and we can rewrite problem (7.34) as

$$\underset{v\in\mathbb{R}^N}{\text{minimize}}\ \Phi(v) + \Psi(Lv). \tag{7.40}$$

Note that in this concise formulation, the functions $\Phi \in \Gamma_0(\mathbb{R}^N)$ and $\Psi \in \Gamma_0(\mathbb{R}^n)$ are nonsmooth. Now let us introduce the functions

$$\boldsymbol{F}\colon (v,s) \mapsto \Phi(v) + \Psi(s) \quad \text{and} \quad \boldsymbol{G} = \iota_{\boldsymbol{V}}, \tag{7.41}$$

where $\boldsymbol{V} = \operatorname{gra} L = \big\{(v,s) \in \mathbb{R}^N \times \mathbb{R}^n \;\big|\; Lv = s\big\}$ is the graph of L, and $\iota_{\boldsymbol{V}}$ is the indicator function of $\boldsymbol{V}$. Using the variable $\boldsymbol{x} = (v,s)$, we see that (7.40) reduces to the problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^{N+n}}{\text{minimize}}\ \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x}) \tag{7.42}$$

involving the sum of two functions in $\Gamma_0(\mathbb{R}^{N+n})$ and which can be solved with the Douglas-Rachford algorithm (Bauschke and Combettes, 2011, Section 27.2). Let $\boldsymbol{x}_0 \in \mathbb{R}^{N+n}$, let $\boldsymbol{y}_0 \in \mathbb{R}^{N+n}$, let $\boldsymbol{z}_0 \in \mathbb{R}^{N+n}$, let $\gamma \in\,]0,+\infty[$, and let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence in $]0,2[$ such that $\inf_{k\in\mathbb{N}} \mu_n > 0$ and $\sup_{k\in\mathbb{N}} \mu_n < 2$. The Douglas-Rachford algorithm

$$\begin{array}{l} \text{for } k = 0,1,\ldots \\ \left\lfloor \begin{array}{l} \boldsymbol{x}_{k+1} = \operatorname{prox}_{\gamma\boldsymbol{G}} \boldsymbol{y}_k \\ \boldsymbol{z}_{k+1} = \operatorname{prox}_{\gamma\boldsymbol{F}} (2\boldsymbol{x}_{k+1} - \boldsymbol{y}_k) \\ \boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \mu_k(\boldsymbol{z}_{k+1} - \boldsymbol{x}_{k+1}) \end{array} \right. \end{array} \tag{7.43}$$

produces a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ which converges to a solution to (7.42) (Bauschke and Combettes, 2011, Corollary 27.4).

We can express the components of the proximity operators $\operatorname{prox}_{\gamma\boldsymbol{F}}$ and $\operatorname{prox}_{\gamma\boldsymbol{G}}$ in terms of the proximity operators of $\operatorname{prox}_{\gamma\psi_i}$ and $\operatorname{prox}_{\gamma f_j}$, for $i = 1,\ldots,n$ and $j = 1,\ldots,m$, and we obtain a block coordinate Rouglas-Rachford algorithm of the general form of (7.43).

Furthermore, rather than evaluating each block i and j at every iteration, following the analysis of Combettes and Pesquet (2015, Corollary 5.5) we can update a randomly chosen subset of the blocks, while still guaranteeing convergence of the algorithm.

We defer the details of the derivation to Appendix F and we obtain the algorithm outlined below. To this end we introduce some additional notation, and set, for every $j \in \{1,\dots,m\}$,

$$R_j = L_j^\top (\operatorname{Id} + LL^\top)^{-1} = L_j^\top \left(\operatorname{Id} + \sum_{j=1}^{m} L_j L_j^\top \right)^{-1}. \tag{7.44}$$

We denote by $x_{j,k} \in \mathbb{R}^{r_j}$ the jth component of x_k, by $v_{j,k} \in \mathbb{R}^{r_j}$ the jth component of v_k, by $z_{j,k} \in \mathbb{R}^{r_j}$ the jth component of z_k. Finally, we denote by $\eta_{i,k} \in \mathbb{R}$ the ith component of y_k, by $\tau_{i,k} \in \mathbb{R}$ the ith component of t_k, and by $\sigma_{i,k} \in \mathbb{R}$ the ith component of s_k. The result follows.

Theorem 100. *Let* $\mathsf{D} = \{0,1\}^{m+n} \smallsetminus \{0\}$*, let* $\gamma \in]0,+\infty[$*, let* $(\mu_k)_{k\in\mathbb{N}}$ *be a sequence in* $]0,2[$ *such that* $\inf_{k\in\mathbb{N}} \mu_n > 0$ *and* $\sup_{k\in\mathbb{N}} \mu_n < 2$*, let* $(v_{j,0})_{1\le j\le m}$ *and* $(x_{j,0})_{1\le j\le m}$ *be in* $\mathbb{R}^N$*, let* $y_0 = (\eta_{i,0})_{1\le i\le n} \in \mathbb{R}^n$*, and let* $(\varepsilon_k)_{k\in\mathbb{N}} = (\varepsilon_{1,k}\dots,\varepsilon_{m+n,k})_{k\in\mathbb{N}}$ *be identically distributed* D*-valued random variables such that, for every* $i \in \{1,\dots,m+n\}$*,* $\operatorname{Prob}[\varepsilon_{i,0} = 1] > 0$*. Iterate*

```
for $k = 0,1,\dots$
    $q_k = \sum_{j=1}^m L_j x_{j,k} - y_k$
    for $j = 1,\dots,m$
        $v_{j,k+1} = v_{j,k} + \varepsilon_{j,k}\big(x_{j,k} - R_j q_k - v_{j,k}\big)$
        $x_{j,k+1} = x_{j,k} + \varepsilon_{j,k}\mu_k\big(\operatorname{prox}_{\gamma\lambda f_j}(2v_{j,k+1} - x_{j,k}) - v_{j,k+1}\big)$        (7.45)
    $s_{k+1} = \sum_{j=1}^m L_j v_{j,k+1}$
    for $i = 1,\dots,n$
        $\eta_{i,k+1} = \eta_{i,k} + \varepsilon_{m+i,k}\mu_k\big(\operatorname{prox}_{\gamma\psi_i}(2\sigma_{i,k+1} - \eta_{i,k}) - \sigma_{i,k+1}\big).$
```

Suppose that the random vectors $(\varepsilon_k)_{k\in\mathbb{N}}$ *and* $(x_k, y_k)_{k\in\mathbb{N}}$ *are independent. Then, for every* $j \in \{1,\dots,m\}$*,* $(v_{j,k})_{k\in\mathbb{N}}$ *converges almost surely to a vector* v_j *and* $w = \sum_{j=1}^m B_j v_j$ *is a solution to* (7.32).

Proof. It follows from Combettes and Pesquet (2015, Corollary 5.5) that $(v_{1,k},\dots,v_{m,k})_{k\in\mathbb{N}}$ converges almost surely to a solution to (7.34). In turn, w solves (7.32). □

Remark 101. *The matrices* $(R_j)_{1\le j\le m}$ *of* (7.44) *are computed off-line only once and they intervene in algorithm* (7.45) *only via matrix-vector multiplications.*

Remark 102. *It follows from the result of Combettes and Pesquet (2015) that, under suitable qualification conditions, the conclusions of Theorem 88 remain true for a general choice of the functions* $f_j \in \Gamma_0(\mathbb{R}^{r_j})$ *and, for every* $\beta \in \mathbb{R}$*,* $\ell_i(\cdot,\beta) \in \Gamma_0(\mathbb{R})$*, and also when* B *does not have full rank. This allows us to solve the more general versions of* (7.33) *in which the regularizer is not a norm but a function of the form* (7.3).

7.4 Numerical Experiments

In this section we present numerical experiments applying the random sweeping stochastic block algorithm outlined in Section 7.3 to sparse problems in which the regularization penalty is a norm fitting our interpolation framework, as described in Assumption 85. The goal of these experiments is to show concrete applications of the class of norms discussed in this chapter and to illustrate the behavior of the proposed random block-iterative proximal splitting algorithm. Let us stress that these appear to be the first numerical experiments on this kind of iteration-convergent, block-coordinate method for completely nonsmooth optimization problems.

The setting we consider is binary classification with the hinge loss and a latent group lasso penalty (Jacob et al., 2009b). Each data matrix $A \in \mathbb{R}^{n \times d}$ is generated with i.i.d. Gaussian entries and each row a_i of A is normalized to have unit ℓ_2 norm. Similarly the weight vector $w \in \mathbb{R}^d$ has i.i.d. Gaussian entries and is normalized after applying a sparsity constraint. The n observations are generated as $\beta_i = \text{sign}(\langle a_i \mid w \rangle)$. A randomly chosen subset of the observations have their sign reversed, with the value of the noise determining the size of the subset, expressed as a percentage of the total observations. For the implementation of the algorithm, we require the proximity operators of the functions f_j and ψ_i in Theorem 88. In this case, these are the ℓ_2-norm and the hinge loss, which have straightforward proximity operators.

In large-scale applications it is impossible to activate all the functions and all the blocks due to computing and memory limitations. Our random sweeping algorithm gives us the possibility of activating only some of them by toggling the activation variables $\varepsilon_{j,k}$ and $\varepsilon_{m+i,k}$. The vectors are updated only when the corresponding activation variable is equal to 1; otherwise it is equal to 0 and no update takes place. In our experiments we always activate the latter as they are associated to the set of training points, which is typically small in sparsity regularization problems. On the other hand, only a fraction α of the former parameters are activated, which is achieved by sampling, at each iteration k, a subset $\lfloor m\alpha \rfloor$ of the components $j \in \{1, \ldots, m\}$ uniformly without replacement. In light of Theorem 88, α can be very small, while still guaranteeing convergence of the iterates. It is natural to ask to what extent these selective updates slow down the algorithm with respect to the hypothetical fully updated version in which sufficient computational power and memory are available.

To investigate this aspect, in our first experiment A is 1000×10000, the true model vector w has sparsity 5%, and we apply 25% classification noise. The relaxation parameter μ is set to 1.99, the proximal parameter γ is set to 0.01, the regularization parameter λ is set to 0.1. Convergence is measured as the percentage of change in iterates, and we use a tolerance of 10^6. As penalty we used the chain latent group lasso, whereby the groups define contiguous

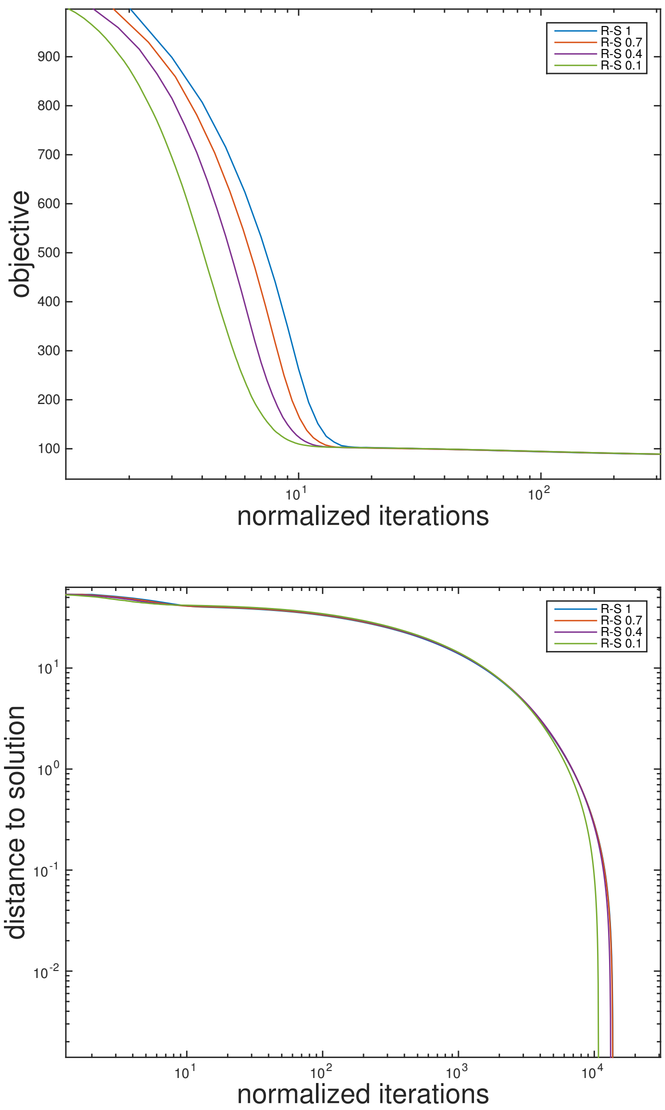


Figure 7.1: Objective for hinge loss classification with the latent group lasso (top), and distance to solution for the same (bottom).

sequences of length 10, with an overlap of length 3. The number of groups is given by the smallest integer no smaller than $\frac{d-\text{overlap}}{\text{length}-\text{overlap}}$, that is 1429. Table 7.1 presents the time, number of iterations, and number of normalized iterations by the activation rate for the hinge loss and latent group lasso penalty for α in $\{0.05, 0.1, 0.2, \ldots, 1.0\}$. Scaling the iterations by the activation rate allows for a fair comparison between regimes since the computational requirements per iteration of the algorithm is proportional to the number of activated blocks. We observe that while the absolute number of iterations increases as the activation rate decreases, scaling these by α reveals that the computational cost is remarkably stable across the different regimes. It follows that using a small value of α does not affect negatively the

Table 7.1: Time, iterations, and normalized iterations to convergence for hinge loss classification with the latent group lasso. $A \in \mathbb{R}^{1000 \times 10000}$, $m = 1429$.

activation rate	time (s)	actual iterations	normalized iterations
1.0	24912	14515	14515
0.9	24517	16047	14443
0.8	26124	18080	14464
0.7	28975	20633	14443
0.6	28223	23872	14323
0.5	29304	28308	14154
0.4	36983	34392	13757
0.3	38100	44080	13224
0.2	50484	62664	12533
0.1	62213	100829	10083
0.05	92368	166111	8306

computational performance, and this can be beneficial when applied to penalties with a large number of blocks. We stress that in large scale problems in which memory space and processing power are limited, the standard optimization algorithm with full activation rate is not suitable (see (F.31) in Appendix F), whereas our random sweeping procedure can be easily implemented. Interestingly, Table 1 indicates that the normalized number of iterations are not affected and in fact slightly improves as the activation rate decreases.

To investigate this further, we present a second experiment using the k-support norm penalty of Argyriou et al. (2012). In this case A is 20×25 and $k = 4$, for α in $\{0.005, 0.001, 0.005, 0.01, 0.05, 0.1, 0.2, \ldots, 1.0\}$. As the number of groups is given by $\binom{d}{k}$, even for this modest setting the number of groups is $m = 12650$, which considerably exceeds d and n. Table 7.2 shows the same metrics as the first experiment. We again observe that performance is consistent over the range of α.

Figure 7.1 depicts (top) the objective values for classification with the hinge loss and a latent group lasso chain penalty, and (bottom) the distance to the limiting solution for various activation rates. The iterations were normalized by the activation rate. We note that the paths are similar for all activation rates, and the convergence is similarly fast for small α. This reinforces our findings that choosing an activation rate lower than 100% does not lead to any deterioration in performance, and in some cases we can even make computational gains.

Finally, for completeness we also consider a standard lasso regression problem (Tibshirani, 1996) with the square loss function and the ℓ_1-norm penalty. In this case we found that, in spite of its greater generality, our random sweeping algorithm remains competitive with FISTA and the forward-backward algorithms, which are both applicable only to the case of differentiable loss functions with Lipschitz gradients. In addition, these experiments indicate that the iterates of our algorithm have a faster convergence rate when the distance to the

Table 7.2: Time, iterations and normalized iterations to convergence for hinge loss classification with the k-support norm. $A \in \mathbb{R}^{20\times 25}$, $k = 4$, $m = 12650$.

activation rate	time (s)	actual iterations	normalized iterations
1.0	388	463	463
0.9	453	610	549
0.8	446	681	544
0.7	526	773	541
0.6	454	894	536
0.5	425	1065	533
0.4	482	1281	512
0.3	389	1694	508
0.2	423	2557	511
0.1	469	4851	485
0.05	1054	9402	470
0.010	1625	41633	416
0.005	1907	77066	385
0.0010	4583	294481	294
0.0005	7369	471137	236

solution is concerned.

We let A be 2000×5000, and set $\lambda = 0.01$ and $\gamma = 1$. The observations $b \in \mathbb{R}^n$ are generated as $b = Aw + r$ where the components of r are i.i.d. Gaussian noise with a standard deviation of 0.001. The number of groups is 5000, and we vary α in $\{0.05, 0.1, 0.2, \ldots, 1.0\}$. Figure 7.2 depicts the objective values, and Table 7.3 depicts running time, iterations and normalized iterations for the random sweeping algorithm. As in the non smooth experiments, we observe the same favorable behaviour relative to the choice of α.

With respect to parameter selection, we found that choosing μ close to 2 generally worked well. Furthermore, in each experimental setting, the random sweeping algorithm is robust to the value of γ, which only affects the speed of convergence. We found that in general choosing γ in $\{0.01, 0.1, 1\}$ gave good performance, in particular when γ and λ were related by only a few orders of magnitude.

As it can be seen from the plots, our method works well for various ranges of activation rates and remains competitive even for small activation rates. Furthermore, our method remains competitive with FISTA and the forward-backward splitting algorithms despite its greater generality and the fact that, unlike the former, it guarantees the convergence of the iterates. What is most remarkable in the experiments is that the iterates of our algorithm have a faster convergence to the solution to the underlying optimization problem. This highlights the fact that in problems in which convergence of the iterates is the main focus, looking at the convergence of the objective values may be misleading.

Table 7.3: Time, iterations and normalized iterations to convergence for regression with square loss and ℓ_1-norm penalty. $A \in \mathbb{R}^{2000 \times 5000}$, $m = 5000$.

activation rate	time (s)	iterations	iterations (normalized)
1.0	5373	1643	1643
0.9	690	215	194
0.8	426	141	113
0.7	349	124	87
0.6	299	125	75
0.5	289	138	69
0.4	320	164	66
0.3	344	213	64
0.2	497	313	63
0.1	674	613	61
0.05	1140	1191	60
0.010	4406	5607	56
0.005	10637	10780	54
0.0010	37566	46108	46

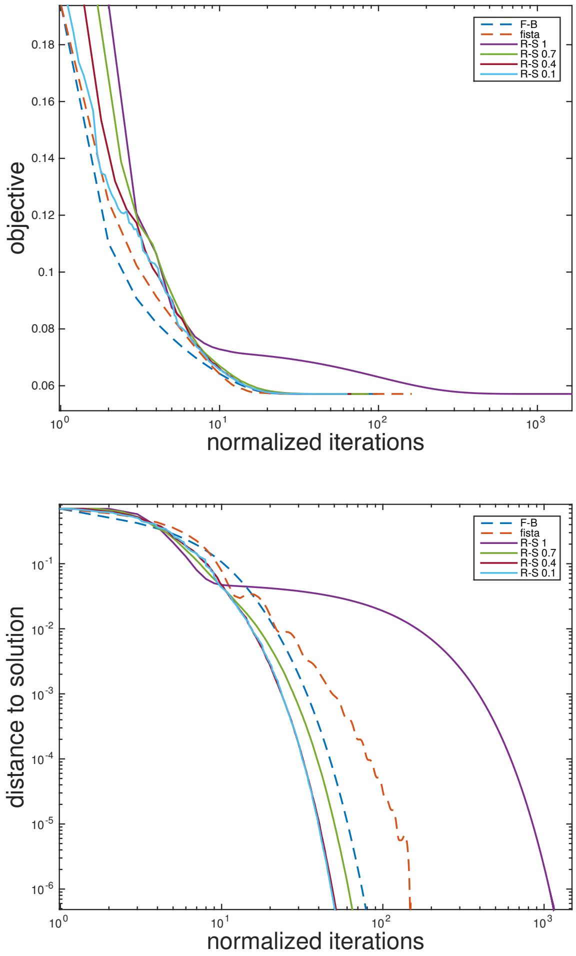


Figure 7.2: Objectives for square loss regression with ℓ_1 penalty (top), and distance to final solution for same (bottom).

Chapter 8

Conclusions

In this thesis we investigated penalties for structured sparsity. In particular we studied norms which are defined by a variational problem. Compared to simple norms such as the ℓ_1-norm, the penalties are more complex, for instance being intractable in the general case, or requiring an iterative algorithm to compute. Nonetheless, the additional structure that these norms provide can lead to improved learning.

We showed that the k-support norm can be written as a so called Θ norm, an infimum of quadratics parameterized over particular a convex set. A natural generalization of the k-support norm leads to the *box*-norm, which we showed is a Moreau envelope of the former. We provided an efficient computation of the proximity operator, which allows large scale regularization problems to be solves using optimal first order numerical algorithms, such as ISTA.

Turning our attention to matrices, we introduced the orthogonally invariant spectral k-support norm and spectral box-norm, each of which is induced by the respective vector norms. We showed that that spectral box-norm is essentially equal to the cluster norm for multitask learning, and we showed how to solve regularization problems with the centered norm. Finally we provided numerical experiments which empirically validate the performance of the norms on real data sets for low rank matrix learning and multitask learning.

We introduced two further extensions of the k-support norm. The (k,p)-support norm allows us to tune the curvature of the sorted underlying components. When applied to the matrix setting, this allows us to tune the spectral decay of the underlying model. A further generalization of the norm is to tensors. In this setting the k-support norm constrains the rank of one of the matricizations of the tensor. In numerical experiments on real low rank matrix and low rank tensor datasets we showed the norms outperforms the respective baseline regularizers.

Finally, we introduced a framework for norms which are defined as an optimal interpolation problem. The k-support norms, along with a number of well known penalties are captured by this general formulation. For the particular instance of norms which are defined via an infimal convolution we provided a random sweeping algorithm which can be used to solve Tikhonov regularization problems including when the loss is non smooth, and we illustrated the sensitivity of the algorithm to the sweeping parameter.

In conclusion, the structured norms that we introduced can lead to improved performance in structured sparsity problems across a variety of domains, and the benefits outweigh the complexity inherent in these penalties.

8.1 Further Work

The material that we presented in this thesis can naturally be extended in a number of directions. For each of the norms discussed our primary focus was the properties of the norm, including its dual norm, its relationship to existing penalties, and optimization considerations. Aside from Rademacher complexities, a key omission from our study are statistical guarantees for the performance of the penalties. Various authors have started to fill the gaps. In particular we highlight Chatterjee et al. (2014); Chen and Banerjee (2016), who provide bounds for the Gaussian width of the k-support norm and spectral k-support norm, Gunasekar et al. (2015), who analyze the consistency of matrix completion with the spectral k-support norm and Richard et al. (2014), who compute the statistical dimension of the vector norm. A similar analysis for the vector and box-norms, and further analysis such as support recovery guarantees would be of interest in future work.

We provided an algorithm to compute the proximity operator for the (k,p)-support norm in the case where p is finite, however in the $p = \infty$ case we were not able to do so, and it would be of interesting to complete the analysis for this case.

The tensor k-support norm was introduced via its unit ball, which is a union of sets of matricizations of bounded rank, which means that one of the matricizations is constrained. It would be interesting to consider the set defined as the intersection of the same sets. This smaller unit ball relates to tensors for which all of the matricizations have a rank restriction. The optimization is challenging to perform in general, however the case of $p = \infty$ lends itself to analysis, using the projection operator of the (k,p)-support norm. We leave the details to a future work.

The Θ-norms form a rich class of penalties, subsuming for example the Λ-norms of Micchelli et al. (2013), and the k-support norm. It can be shown that the (k,p)-support norm is included in this class when $p \in [1,2]$, but it is not known whether this holds for $p \in [2,\infty)$; indeed we conjecture that this is not the case. As we outlined, the generalization to the Θ p-norms allows us to naturally include the (k,p)-norms in the class. An interesting point for further research is whether, given an arbitrary norm, it is possible to conclude whether this norm can be expressed as a primal or dual Θ-norm or a Θ p-norm, and how to determine the parameter set Θ in the affirmative cases.

We introduced a very general optimal interpolation framework in Chapter 7. The algorithm that we provided caters only to the case where the principle constituent norm is the ℓ_1-norm, that is the objective is evaluated as a sum. The case of the max (ℓ_∞-norm) would be of interest to study, as would more general formulations. The framework further lends itself to defining new norms, simply by choosing the constitutent components appropriately. In future work it would be of interest to introduce new regularization penalties which are motivated by machine learning problems, and show that their performance improves upon existing penalties.

Appendix A

Appendix: Background

In this appendix we present derivations for Chapter 2. For completeness we restate each result throughout the appendices.

Proposition 3 (Hölder Inequality)**.** *Let* $x, y \in R^d$*, and let* $p, q \in (1, \infty)$ *such that* $\frac{1}{p} + \frac{1}{q} = 1$*. Then*

$$\langle x \mid y \rangle \leq \|x\|_p \|y\|_q, \tag{A.1}$$

and equality holds when $w_i = \alpha sign(y_i)(|y_i|)^{\frac{q}{p}}$*, for any* $\alpha \in \mathbb{R}_+$*. Similarly*

$$\langle x \mid y \rangle \leq \|x\|_1 \|y\|_\infty, \tag{A.2}$$

and equality holds when

$$x_i = \begin{cases} \alpha sign(y_i) & if\ i \in \mathcal{I} \\ 0 & if\ i \notin \mathcal{I} \end{cases}, \tag{A.3}$$

where $\mathcal{I} = \left\{ i = 1 \ldots n \;\middle|\; i = \operatorname{argmax}_{i=1}^d |y_i| \right\}$*, for any* $\alpha \in \mathbb{R}_+$*.*

Proof. The inequalities follow by duality, so it remains to prove the two cases of equality. For $p, q \in (1, \infty)$, substituting the proposed value for x_i the left hand side becomes

$$\sum_{i=1}^d x_i y_i = \sum_{i=1}^d \alpha \text{sign}(y_i)(|y_i|)^{\frac{q}{p}} y_i = \alpha \sum_{i=1}^d (|y_i|)^{\frac{q}{p}+1} = \alpha \sum_{i=1}^d (|y_i|)^q = \alpha \|y\|_q^q, \tag{A.4}$$

using the fact that p and q are Hölder conjugates. The right hand side becomes

$$\|x\|_p \|y\|_q = \left(\sum_{i=1}^d (|\alpha \text{sign}(y_i)(|y_i|)^{\frac{q}{p}}|)^p \right)^{\frac{1}{p}} \left(\sum_{i=1}^d (|y_i|)^q \right)^{\frac{1}{q}} = \alpha \left(\sum_{i=1}^d (|y_i|)^q \right)^{\frac{1}{p}+\frac{1}{q}} = \alpha \|y\|_q^q, \tag{A.5}$$

as required. For the second equality, substituting the corresponding value for x_i the left hand side becomes

$$\sum_{i=1}^{d} x_i y_i = \alpha \sum_{i \in \mathcal{I}} \text{sign}(y_i) y_i = \alpha c |y_m|, \tag{A.6}$$

where $m \in \mathcal{I}$ hence y_m is the largest component of y in absolute value and $c = |\mathcal{I}|$. The right hand side becomes

$$\|x\|_1 \|y\|_\infty = \left(\sum_{i=1}^{d} |x_i| \right) |y_m| = \alpha \left(\sum_{i \in \mathcal{I}} 1 \right) |y_m| = \alpha c |y_m|, \tag{A.7}$$

as required. See also Marshall and Olkin (1979, Ch. 16 Sec. D.1). □

Proposition 6. *Let $\|\cdot\|$ be an orthogonally invariant norm induced by the symmetric gauge function g. Then, for $Y \in \mathbb{R}^{d \times m}$ the dual norm is given by $\|Y\|_* = g_*(\sigma(Y))$ where g_* is the dual of g.*

Proof. Let Y have singular value decomposition $Y = A\Sigma_Y B^\top$. Applying the definition of the dual norm,

$$\begin{aligned}
\|Y\|_* &= \sup\left\{\langle Y \mid X\rangle \;\middle|\; \|X\| \le 1\right\} && \text{(A.8)}\\
&= \sup\left\{\operatorname{tr}(Y^\top X) \;\middle|\; \|X\| \le 1\right\} && \text{(A.9)}\\
&= \sup\left\{\operatorname{tr}(B\Sigma_Y^\top A^\top A\Sigma_X B^\top) \;\middle|\; \|\Sigma_X\| \le 1\right\} && \text{(A.10)}\\
&= \sup\left\{\operatorname{tr}(\Sigma_Y \Sigma_X) \;\middle|\; g(\sigma(X)) \le 1\right\} && \text{(A.11)}\\
&= \sup\left\{\langle \sigma(Y) \mid \sigma(X)\rangle \;\middle|\; g(\sigma(X)) \le 1\right\} && \text{(A.12)}\\
&= g_*(\sigma(Y)), && \text{(A.13)}
\end{aligned}$$

where Σ_X is the matrix of singular values of X, we have used the facts that $\|\cdot\|$ is orthogonally invariant, Σ_Y is diagonal, and we may choose X to have the same ordered system of left and right singular values as Y. □

Lemma 36. *Let $C \subset X$ be a bounded, convex, balanced, and absorbing set. The Minkowski functional μ_C of C is a norm on X.*

Proof. We give a direct proof that μ_C satisfies the properties of a norm. Clearly $\mu_C(w) \ge 0$ for all w, and $\mu_C(0) = 0$. Moreover, as C is bounded, $\mu_C(w) > 0$ whenever $w \ne 0$.

Next we show that μ_C is one-homogeneous. For every $\alpha \in \mathbb{R}$, $\alpha \ne 0$, let $\sigma = \text{sign}(\alpha)$ and note that

$$\begin{aligned}\mu_C(\alpha w) &= \inf\Big\{\lambda > 0 \;\Big|\; \frac{1}{\lambda}\alpha w \in C\Big\} \\ &= \inf\Big\{\lambda > 0 \;\Big|\; \frac{|\alpha|}{\lambda}\sigma w \in C\Big\} \\ &= |\alpha| \inf\Big\{\lambda > 0 \;\Big|\; \frac{1}{\lambda} w \in \sigma C\Big\} \\ &= |\alpha| \inf\Big\{\lambda > 0 \;\Big|\; \frac{1}{\lambda} w \in C\Big\} = |\alpha|\mu_C(w),\end{aligned}$$

where we have made a change of variable and used the fact that $\sigma C = C$.

Finally, we prove the triangle inequality. For every $v, w \in X$, if $v/\lambda \in C$ and $w/\mu \in C$ then setting $\gamma = \lambda/(\lambda+\mu)$, we have

$$\frac{v+w}{\lambda+\mu} = \gamma\frac{v}{\lambda} + (1-\gamma)\frac{w}{\mu}$$

and since C is convex, then $\frac{v+w}{\lambda+\mu} \in C$. We conclude that $\mu_C(v+w) \le \mu_C(v) + \mu_C(w)$. The proof is completed. □

Example 38. *Examples of conjugates.*

a) *Let $C \subset \mathcal{X}$ and let $f = \delta_C$. Then $f^* = \sigma_C$, that is, the conjugate is the support function.*

b) *Let f be a norm $\|\cdot\|$. Then $f^* = \delta_{\mathcal{B}_*}$, where $\mathcal{B}_* = \big\{u \in X \;\big|\; \|u\|_* \le 1\big\}$, that is, the conjugate is the characteristic function of the unit ball of the dual norm.*

c) *Let $f = \frac{1}{2}\|\cdot\|^2$. Then $f^* = \frac{1}{2}\|\cdot\|_*^2$.*

Proof. *a*) By direct computation

$$\begin{aligned}f^*(u) &= \sup_{x\in X}\langle u \mid x\rangle - \delta_C(x) = \sup_{x\in C}\langle u \mid x\rangle \\ &= \sigma_C(u).\end{aligned}$$

b) We have

$$\begin{aligned}f^*(u) &= \sup_{x\in X}\langle u \mid x\rangle - \|x\| = \sup_{x\in X}\|x\|\big(\|u\|_* - 1\big) \\ &= \begin{cases} 0 & \text{if } \|u\|_* \le 1 \\ \infty & \text{if } \|u\|_* > 1 \end{cases} \\ &= \delta_{\mathcal{B}_*}(u),\end{aligned}$$

where we have used Hölder's inequality (Proposition 2).

c) We have

$$\begin{aligned} f^*(u) &= \sup_{x\in X} \langle u \mid x\rangle - \|x\| \\ &= \sup_{x\in X} \|x\|\big(\|u\|_* - \frac{1}{2}\|x\|\big), \end{aligned}$$

again using Hölder's inequality. Writing $t = \|x\|$, the quantity $t(\|u\|_* - 2t)$ is a concave quadratic in $t \in R$, with roots at $t \in \{0, 2\|u\|_*\}$, hence it attains its maximum when $t = \|u\|_*$. Substituting this value we get the desired result.

□

Lemma 31. *A norm is convex and proper.*

Proof. Let $x, y \in X$, $\lambda \in [0,1]$ and let $\|\cdot\|$ be a norm on X. Then

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = \lambda\|x\| + (1-\lambda)\|y\|, \tag{A.14}$$

hence $\|\cdot\|$ is convex. By positivity the norm is bounded below, hence it is also proper. □

Lemma 32. *Let $f : X \to \mathbb{R}$ be positive, homogeneous and convex. Then f satisfies the triangle inequality. If f moreover is only 0 at the origin, then f is a norm.*

Proof. Let $x, y \in X$. By convexity for any $\lambda \in [0,1]$ we have

$$2f(\lambda x + (1-\lambda)y) \leq 2\lambda f(x) + 2(1-\lambda)f(y),$$

hence choosing $\lambda = \frac{1}{2}$ and using homogeneity we get

$$f(x+y) \leq f(x) + f(y),$$

as required, hence f satisfies all the requirements to be a norm, as required. □

Lemma 34. *Let $f, g : X \to \mathbb{R}$ be convex. Then $f + g$ is convex.*

Proof. Let $\lambda \in [0,1]$, and let $x, y \in X$. Then

$$\begin{aligned} (f+g)(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &\leq \lambda f(x) + (1-\lambda)f(y) + \lambda g(x) + (1-\lambda)g(x) \\ &= \lambda(f+g)(x) + (1-\lambda)(f+g)(y). \end{aligned}$$

□

Note that if either of f or g is strictly convex, then the inequality becomes strict.

Lemma 49 (Best rank k matrix approximation)**.** *Let $X \in \mathbb{R}^{d\times m}$ have singular value decomposition $X = U\operatorname{diag}(\sigma(X))V^\top$ and let $k \le q = \min(d,m)$. Then the closest rank k matrix to X in Frobenius norm is given by the matrix $Z = U\operatorname{diag}(\sigma(Z))V^\top$, where $\sigma(Z) = (\sigma_1(X),\ldots,\sigma_k(X),0,\ldots,0)$.*

Proof. We solve the problem

$$\min\left\{\|X-Z\|_{\text{fro}} \,\middle|\, Z\in\mathbb{R}^{d\times m}, \operatorname{rank}(Z)\le k\right\}. \tag{A.15}$$

Using the orthogonal invariance of the Frobenius norm, this is equivalent to

$$\min\left\{\|\operatorname{diag}(\sigma(X)) - U^\top Z V\|_{\text{fro}} \,\middle|\, Z\in\mathbb{R}^{d\times m}, \operatorname{rank}(Z)\le k\right\}. \tag{A.16}$$

As the first term in the norm is diagonal, we minimize the expression by choosing the second term to be diagonal as well, that is by choosing $Z = U\operatorname{diag}(\sigma(Z))V^\top$, for some vector $\sigma(Z)$, hence the problem reduces to

$$\min\left\{\|\sigma(X)-\sigma(Z)\|_2 \,\middle|\, \sigma(Z)\in\mathbb{R}^q, \operatorname{card}(\sigma(Z))\le k\right\}, \tag{A.17}$$

that is

$$\min\left\{\sum_{i=1}^{q}(\sigma_i(X)-\sigma_i(Z))^2 \,\middle|\, \sigma(Z)\in\mathbb{R}^q, \operatorname{card}(\sigma(Z))\le k\right\}. \tag{A.18}$$

As the singular values of X are ordered non increasing by convention, and we are free to choose at most k non negative singular values for Z, the solution is given by identifying the k largest singular values of Z and X, and setting the remainder to zero, as required. □

Proposition 48. *Let $\mathcal{H}$ be a finite dimensional Euclidean space, let $\mathcal{L}$ be continuous, convex, proper and coercive on $\mathcal{H}$ and let Ω be a norm on $\mathcal{H}$. Let $\lambda,\alpha,\beta\in\mathbb{R}_+$. In each of the problems* (2.1)*,* (2.2) *and* (2.3) *the set of solutions is nonempty and compact.*

Proof. For each setting we consider the extended problem

$$\tilde f(x) = \begin{cases} f(x) & \text{if } x\in X \\ \infty & \text{otherwise.}\end{cases} \tag{A.19}$$

Proposition 47 then provides that the set of minima of f over X is nonempty and compact if X is closed, f is lower semicontinuous at each $x\in X$ (hence $\tilde f$ is closed by Proposition 26), and one of the following conditions holds:

i) X is bounded.

ii) Some set $\{x \in X \mid f(x) \leq \gamma\}$ is nonempty and bounded.

iii) $\tilde{f}$ is coercive.

For the optimization problems in question we have as follows

a) For Tikhonov regularization (2.1), X is the entire Euclidean space, and $f = \mathcal{L} + \lambda\Omega$.

b) For Ivanov regularization (2.2), $X = \{x \mid \Omega(x) \leq \beta\}$, and $f = \mathcal{L}$.

c) For Morozov regularization (2.3), $X = \{x \mid \mathcal{L}(x) \leq \gamma\}$, and $f = \Omega$.

In each case X is closed, f is continous, hence lower semicontinuous, and $\tilde{f}$ is coercive, hence Proposition 47 applies as claimed. □

Appendix B

Appendix: The Θ-norm

In this appendix we present derivations for Chapter 3.

Lemma 52. *For $p \in [1,\infty]$ we have*

$$\|w\|_p = \begin{cases} \|w\|_\Theta, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \,\middle|\, \sum_{i=1}^d \theta_i^{\frac{p}{2-p}} \le 1\right\}, \quad \textit{for } p \in [1,2), \\ \|w\|_\Theta = \|w\|_{\Theta,*}, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \,\middle|\, 0 < \theta_i \le 1\right\}, \quad \textit{for } p = 2, \\ \|w\|_{\Theta,*}, & \Theta = \left\{\theta \in \mathbb{R}^d_{++} \,\middle|\, \sum_{i=1}^d \theta_i^{\frac{p}{p-2}} \le 1\right\}, \quad \textit{for } p \in (2,\infty]. \end{cases}$$

Note that the case $p = 2$ follows from the cases $p \in [1,2)$ and $p \in (2,\infty]$ by taking the limit as p tends to 2 from below or above, respectively.

Proof. Consider the case $p \in [1,2)$, define $a = \frac{2}{p}$ and $b = \frac{2}{2-p}$ and note that $a,b \in [1,\infty)$, and $\frac{1}{a} + \frac{1}{b} = 1$. We have for any $\theta \in \Theta$

$$(\|w\|_p)^p = \sum_{i=1}^d \frac{|w_i|^p}{\theta_i^{\frac{p}{2}}} \theta_i^{\frac{p}{2}} \le \left(\sum_{i=1}^d \frac{|x_i|^2}{\theta_i}\right)^{\frac{p}{2}} \left(\sum_{i=1}^d \theta_i^{\frac{p}{2-p}}\right)^{\frac{2-p}{2}} \le \left(\sum_{i=1}^d \frac{|x_i|^2}{\theta_i}\right)^{\frac{p}{2}}, \tag{B.1}$$

hence

$$\|w\|_p \le \inf_{\theta \in \Theta} \left(\sum_{i=1}^d \frac{|x_i|^2}{\theta_i}\right)^{\frac{1}{2}}. \tag{B.2}$$

Now let $\tilde{\theta} = \frac{|w_i|^{2-p}}{\left(\sum_{j=1}^d |w_j|^p\right)^{\frac{2-p}{p}}}$, and note that $\sum_{i=1}^d \tilde{\theta}_i^{\frac{p}{2-p}} = 1$. Furthermore for this choice we have equality, as

$$\left(\sum_{i=1}^{d}\frac{|w_i|^2}{\tilde{\theta}_i}\right)^{\frac{1}{2}} = \left(\sum_{i=1}^{d}|w_i|^p\left(\sum_{j=1}^{d}|w_j|^p\right)^{\frac{2-p}{p}}\right)^{\frac{1}{2}} = \left(\left(\sum_{j=1}^{d}|w_j|^p\right)^{\frac{p+2-p}{p}}\right)^{\frac{1}{2}} \tag{B.3}$$

$$= \left(\sum_{j=1}^{d}|w_j|^p\right)^{\frac{1}{p}} = \|w\|_p. \tag{B.4}$$

It follows that $\|w\|_p = \|w\|_\Theta$. For the case $p \in (2,\infty]$, note that the Hölder conjugate $q = \frac{p}{p-1}$ lies in $[1,2)$, and we have $\frac{p}{2-p} = \frac{q}{q-2}$. By the first part of the proof it follows that

$$\|w\|_p = \|w\|_{q,*} = \|w\|_{\Theta,*}. \tag{B.5}$$

The case $p = 2$ follows from both cases by continuity. □

For Proposition 55 we recall the seminorm:

$$\|w\|_\gamma = \sqrt{\sum_{i:\gamma_i>0}\frac{w_i^2}{\gamma_i}}.$$

Proposition 55. *Let $\gamma^1,\ldots,\gamma^m \in \mathbb{R}^d_+$ such that $\sum_{\ell=1}^m \gamma^\ell \in \mathbb{R}^d_{++}$ and let Θ be defined by $\Theta = \{\theta \in \mathbb{R}^d_{++} \mid \theta = \sum_{\ell=1}^m \lambda_\ell \gamma^\ell,\ \lambda \in \Delta^{m-1}\}$. Then we have, for every $w \in \mathbb{R}^d$, that*

$$\|w\|_\Theta = \inf\left\{\sum_{\ell=1}^{m}\|v_\ell\|_{\gamma^\ell} \,\middle|\, v_\ell \in \mathbb{R}^d,\ \mathrm{supp}(v_\ell) \subset \mathrm{supp}(\gamma^\ell),\ \ell \in \mathbb{N}_m,\ \sum_{\ell=1}^{m} v_\ell = w\right\}. \tag{B.6}$$

Moreover, the unit ball of the norm is given by the convex hull of the set

$$\bigcup_{\ell=1}^{m}\left\{w \in \mathbb{R}^d \mid \mathrm{supp}(w) \subset \mathrm{supp}(\gamma^\ell), \|w\|_{\gamma^\ell} \le 1\right\}. \tag{B.7}$$

Proof. Let $A_\ell = \{w \in \mathbb{R}^d \mid \|w\|_{\gamma^\ell} \le 1,\ \mathrm{supp}(w) \subset \mathrm{supp}(\gamma^\ell)\}$, and define

$$C = \mathrm{co}\bigcup_{\ell=1}^{m} A_\ell.$$

Note that C is bounded and balanced, since each set A_ℓ is so. Furthermore, the hypothesis that $\sum_{\ell=1}^m \gamma^\ell > 0$ ensures that C is absorbing. Hence, by Lemma 36 the Minkowski functional μ_C defines a norm. We rewrite $\mu_C(w)$ as

$$\mu_C(w) = \inf\left\{\lambda \,\middle|\, \lambda > 0,\ w = \lambda\sum_{\ell=1}^{m}\alpha_\ell z_\ell,\ z_\ell \in A_\ell,\ \alpha \in \Delta^{m-1}\right\}$$

where the infimum is over λ, the vectors $z_\ell \in \mathbb{R}^d$ and the vector $\alpha = (\alpha_1,\ldots,\alpha_m)$, and recall Δ^{m-1} denotes the unit simplex in $\mathbb{R}^m$.

The rest of the proof is structured as follows. We first show that $\mu_C(w)$ coincides with the right hand side of equation (B.6), which we denote by $\nu(w)$. Then we show that $\|w\|_\Theta = \mu_C(w)$ by observing that both norms have the same dual norm.

Choose any vectors $v_1,\ldots,v_m \in \mathbb{R}^d$ which satisfies the constraint set in the right hand side of (B.6) and set $\alpha_\ell = \|v_\ell\|_{\gamma^\ell}/(\sum_{k=1}^m \|v_k\|_{\gamma^k})$ and $z_\ell = v_\ell/\|v_\ell\|_{\gamma^\ell}$. We have

$$w = \sum_{\ell=1}^m v_\ell = \Big(\sum_{k=1}^m \|v_k\|_{\gamma^k}\Big)\sum_{\ell=1}^m \alpha_\ell z_\ell.$$

This implies that $\mu_C(w) \le \nu(w)$. Conversely, if $w = \lambda\sum_{\ell=1}^m \alpha_\ell z_\ell$ for some $z_\ell \in A_\ell$ and $\alpha \in \Delta^{m-1}$, then letting $v_\ell = \lambda\alpha_\ell z_\ell$ we have

$$\sum_{\ell=1}^m \|v_\ell\|_{\gamma^\ell} = \sum_{\ell=1}^m \|\lambda\alpha_\ell z_\ell\|_{\gamma^\ell} = \lambda\sum_{\ell=1}^m \alpha_\ell\|z_\ell\|_{\gamma^\ell} \le \lambda.$$

Next, we show that both norms have the same dual norm. When Θ is the interior of $\mathrm{co}\{\gamma^1,\ldots,\gamma^m\}$, the dual norm of $\|\cdot\|_\Theta$, described in Proposition 51, can be written as

$$\|u\|_{\Theta,*} = \sup_{\theta\in\Theta}\sqrt{\sum_{i=1}^d \theta_i u_i^2} = \max_{\ell=1}^m \sqrt{\sum_{i=1}^d \gamma_i^\ell u_i^2}.$$

We now compute the dual of the norm μ_C,

$$\max_{w\in C}\langle w \mid u\rangle = \max\big\{\langle w \mid u\rangle \;\big|\; w \in \cup_{\ell=1}^m A_\ell\big\} = \max_{\ell=1}^m \max_{w\in A_\ell}\langle w \mid u\rangle = \max_{\ell=1}^m \sqrt{\sum_{i=1}^d \gamma_i^\ell u_i^2}. \tag{B.8}$$

It follows that the norms share the same dual norm, hence $\mu_C(\cdot)$ coincides with $\|\cdot\|_\Theta$. □

Theorem 63. *For every $w \in \mathbb{R}^d$ it holds that*

$$\|w\|_{\mathrm{box}}^2 = \frac{1}{b}\|w_Q\|_2^2 + \frac{1}{p}\|w_I\|_1^2 + \frac{1}{a}\|w_L\|_2^2,$$

where $w_Q = (|w|_1^\downarrow,\ldots,|w|_q^\downarrow)$, $w_I = (|w|_{q+1}^\downarrow,\ldots,|w|_{d-\ell}^\downarrow)$, $w_L = (|w|_{d-\ell+1}^\downarrow,\ldots,|w|_d^\downarrow)$, q and ℓ are the unique integers in $\{0,\ldots,d\}$ that satisfy $q+\ell \le d$,

$$\frac{|w_q|}{b} \ge \frac{1}{p}\sum_{i=q+1}^{d-\ell}|w_i| > \frac{|w_{q+1}|}{b}, \qquad \frac{|w_{d-\ell}|}{a} \ge \frac{1}{p}\sum_{i=q+1}^{d-\ell}|w_i| > \frac{|w_{d-\ell+1}|}{a},$$

$p = c - qb - \ell a$ and we have defined $|w_0| = \infty$ and $|w_{d+1}| = 0$. Furthermore, the minimizer θ

has the form

$$\theta_i = \begin{cases} b, & \text{if } i \in \{1,\dots,q\}, \\ p\frac{|w_i|}{\sum_{j=q+1}^{d-\ell}|w_j|}, & \text{if } i \in \{q+1,\dots,d-\ell\}, \\ a, & \text{otherwise}. \end{cases}$$

Proof. We solve the constrained optimization problem

$$\inf\left\{\sum_{i=1}^{d}\frac{w_i^2}{\theta_i} \;\middle|\; a \le \theta_i \le b, \sum_{i=1}^{d}\theta_i \le c\right\}. \tag{B.9}$$

To simplify the notation we assume without loss of generality that w_i are positive and ordered nonincreasing, and note that the optimal θ_i are ordered non increasing. To see this, let $\theta^* = \operatorname{argmin}_{\theta\in\Theta}\sum_{i=1}^{d}\frac{w_i^2}{\theta_i}$. Now suppose that $\theta_i^* < \theta_j^*$ for some $i < j$ and define $\hat{\theta}$ to be identical to θ^*, except with the i and j elements exchanged. The difference in objective values is

$$\sum_{i=1}^{d}\frac{w_i^2}{\hat{\theta}_i} - \sum_{i=1}^{d}\frac{w_i^2}{\theta_i^*} = (w_i^2 - w_j^2)\left(\frac{1}{\theta_j^*} - \frac{1}{\theta_i^*}\right),$$

which is negative so θ^* cannot be a minimizer.

We further assume without loss of generality that $w_i \neq 0$ for all i, and $c \le db$ (see Remark 63a below). The objective is continuous and we take the infimum over a closed bounded set, so a solution exists and it is unique by strict convexity. Furthermore, since $c \le db$, the sum constraint will be tight at the optimum. Consider the Lagrangian function

$$L(\theta,\alpha) = \sum_{i=1}^{d}\frac{w_i^2}{\theta_i} + \frac{1}{\alpha^2}\left(\sum_{i=1}^{d}\theta_i - c\right), \tag{B.10}$$

where $1/\alpha^2$ is a strictly positive multiplier, and α is to be chosen to make the sum constraint tight, call this value α^*. Let θ^* be the minimizer of $L(\theta,\alpha^*)$ over θ subject to $a \le \theta_i \le b$.

We claim that θ^* solves equation (B.9). Indeed, for any $\theta \in [a,b]^d$, $L(\theta^*,\alpha^*) \le L(\theta,\alpha^*)$, which implies that

$$\sum_{i=1}^{d}\frac{w_i^2}{\theta_i^*} \le \sum_{i=1}^{d}\frac{w_i^2}{\theta_i} + \frac{1}{(\alpha^*)^2}\left(\sum_{i=1}^{d}\theta_i - c\right).$$

If in addition we impose the constraint $\sum_{i=1}^{d}\theta_i \le c$, the second term on the right hand side is at most zero, so we have for all such θ that

$$\sum_{i=1}^{d} \frac{w_i^2}{\theta_i^*} \leq \sum_{i=1}^{d} \frac{w_i^2}{\theta_i},$$

whence it follows that θ^* is the minimizer of (B.9).

We can therefore solve the original problem by minimizing the Lagrangian (B.10) over the box constraint. Due to the coupling effect of the multiplier, the problem is separable, and we can solve the simplified problem componentwise (see Micchelli et al., 2013, Theorem 3.1). For completeness we repeat the argument here. For every $w_i \in \mathbb{R}$ and $\alpha > 0$, the unique solution to the problem $\min\left\{\frac{w_i^2}{\theta} + \frac{\theta}{\alpha^2} \;\middle|\; a \leq \theta \leq b\right\}$ is given by

$$\theta = \begin{cases} b, & \text{if } \alpha|w_i| > b, \\ \alpha|w|, & \text{if } b \geq \alpha|w_i| \geq a, \\ a, & \text{if } a > \alpha|w_i|. \end{cases} \tag{B.11}$$

Indeed, for fixed w_i, the objective function is strictly convex on $\mathbb{R}^d_{++}$ and has a unique minimum on $(0,\infty)$ (see Figure 1.b in Micchelli et al. (2013) for an illustration). The derivative of the objective function is zero for $\theta = \theta^* := \alpha|w_i|$, strictly positive below θ^* and strictly increasing above θ^*. Considering these three cases the result follows and θ is determined by (B.11) where α satisfies $\sum_{i=1}^{d} \theta_i(\alpha) = c$.

The minimizer then has the form

$$\theta = (\underbrace{b,\ldots,b}_{q}, \theta_{q+1},\ldots,\theta_{d-\ell}, \underbrace{a,\ldots,a}_{\ell}),$$

where $q, \ell \in \{0,\ldots,d\}$ are determined by the value of α which satisfies

$$S(\alpha) = \sum_{i=1}^{d} \theta_i(\alpha) = qb + \sum_{i=q+1}^{d-\ell} \alpha|w_i| + \ell a = c,$$

i.e. $\alpha = p/\left(\sum_{i=q+1}^{d-\ell} |w_i|\right)$, where $p = c - qb - \ell a$.

The value of the norm follows by substituting θ into the objective and we get

$$\|w\|_{\text{box}}^2 = \sum_{i=1}^{q} \frac{|w_i|^2}{b} + \frac{1}{p}\Big(\sum_{i=q+1}^{d-\ell} |w_i|\Big)^2 + \sum_{i=d-\ell+1}^{d} \frac{|w_i|^2}{a} = \frac{1}{b}\|w_Q\|_2^2 + \frac{1}{p}\|w_I\|_1^2 + \frac{1}{a}\|w_L\|_2^2,$$

as required. We can further characterize q and ℓ by considering the form of θ. By construction we have $\theta_q \geq b > \theta_{q+1}$ and $\theta_{d-\ell} > a \geq \theta_{d-\ell+1}$, or equivalently

$$\frac{|w_q|}{b} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{q+1}|}{b} \quad \text{and} \quad \frac{|w_{d-\ell}|}{a} \geq \frac{1}{p} \sum_{i=q+1}^{d-\ell} |w_i| > \frac{|w_{d-\ell+1}|}{a}.$$

The proof is completed. □

Remark 63a. *The case where some w_i are zero follows from the case that we have considered in the theorem. If $w_i = 0$ for $n < i \leq d$, then clearly we must have $\theta_i = a$ for all such i. We then consider the n-dimensional problem of finding $(\theta_1, \ldots, \theta_n)$ that minimizes $\sum_{i=1}^n \frac{w_i^2}{\theta_i}$, subject to $a \leq \theta_i \leq b$, and $\sum_{i=1}^n \theta_i \leq c'$, where $c' = c - (d-n)a$. As $c \geq da$ by assumption, we also have $c' \geq na$, so a solution exists to the n-dimensional problem. If $c' \geq bn$, then a solution is trivially given by $\theta_i = b$ for all $i = 1, \ldots, n$. In general, $c' < bn$, and we proceed as per the proof of the theorem. Finally, a vector that solves the original d-dimensional problem will be given by $(\theta_1, \ldots, \theta_n, a, \ldots, a)$.*

Proposition 68. *Let $\Theta \subset \mathbb{R}^d_{++}$ be convex and bounded, and let $p \in [1, \infty]$ and let $q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$. Define the following for $w, u \in \mathbb{R}^d$*

$$\|w\| = \inf_{\theta \in \Theta} \phi_p^{-1}\left(\sum_{i=1}^d \theta_i \phi_p\left(\frac{|w_i|}{\theta_i}\right)\right), \tag{B.12}$$

$$\|u\|_* = \sup_{\theta \in \Theta} \phi_q^{-1}\left(\sum_{i=1}^d \theta_i \phi_q(|u_i|)\right). \tag{B.13}$$

Then the expressions define norms, and furthermore they are dual to each other.

In order to prove Proposition 68 we require the following auxiliary result, see e.g. Louditski (2016, Remark 3.2.1).

Lemma 68a. *If $f(x, y) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is jointly convex, and the function*

$$g(x) = \inf_y f(x, y)$$

is proper, then g is convex.

We further require the following result, see e.g. Borwein and Vanderwerff (2010, Th. 5.4.6.).

Lemma 68b. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^d$, with dual norm $\|\cdot\|_*$. Let $h(w) = \frac{1}{p}\|w\|^p$. Then the Fenchel conjugate h^* satisfies $h^*(u) = \frac{1}{q}\|u\|_*^q$, where $\frac{1}{p} + \frac{1}{q} = 1$.*

The proof of Proposition 68 follows, using the same proof method as in Proposition 51.

Proof. First note that the objective in (3.29) is positive, homogeneous and nondegenerate. Furthermore, it is proper as it is nowhere $-\infty$, and for example, it takes the value zero at the origin. It is jointly convex in w and θ, hence by Lemma 68a, Equation B.12 is convex in w. By Lemma 32 it also satisfies the triangle inequality, hence it is a norm. Conversely expression (3.30) defines a norm as it is a supremum of norms.

Let $q > 1$. Define $h^*(u) = \frac{1}{q}\|u\|_*^q$, for some $q > 1$, and let p be such that $\frac{1}{p} + \frac{1}{q} = 1$. By Lemma 68b we have, for $w \in \mathbb{R}^d$, $h^{**}(w) = \frac{1}{p}\|w\|^p$. Moreover, as h^* is convex, proper and zero at the origin, $h^{**} = h$. We compute h explicitly for $w \in \mathbb{R}^d$:

$$h(w) = \sup_u \langle w \mid u \rangle - f^*(u) = \sup_u \sum_{i=1}^d w_i u_i - \sup_\theta \sum_{i=1}^d \frac{1}{q}\theta_i |u_i|^q = \sup_u \inf_\theta \sum_{i=1}^d w_i u_i - \frac{1}{q}\theta_i |u_i|^q,$$

where we have used the relation $-\max f = \min(-f)$. As in Proposition 51, this is a minimax problem in the sense of von Neumann (Prop. 2.6.3 in Bertsekas et al., 2003), and we can exchange the order of the $\inf$ and the $\sup$, and solve the latter componentwise. The gradient with respect to u_i is $w_i - \theta_i |u_i|^{q-1}\operatorname{sign}(u_i)$, which is zero when $|u_i|^{q-1} = \frac{w_i}{\theta_i}\operatorname{sign}(u_i)$. We therefore require $\operatorname{sign}(u_i) = \operatorname{sign}(w_i)$ and the condition becomes $u_i = \operatorname{sign}(w_i)\left(\frac{|w_i|}{\theta_i}\right)^{\frac{1}{q-1}}$. Substituting we obtain

$$h(w) = \inf_\theta \sum_{i=1}^d w_i \operatorname{sign}(w_i)\left(\frac{|w_i|}{\theta_i}\right)^{\frac{1}{q-1}} - \frac{1}{q}\theta_i \left(\frac{|w_i|}{\theta_i}\right)^{\frac{q}{q-1}} \tag{B.14}$$

$$= \inf_\theta \sum_{i=1}^d \left(\frac{|w_i|}{\theta_i}\right)^{\frac{1}{q-1}} \left(|w_i| - \left(1 - \frac{1}{p}\right)\theta_i \frac{|w_i|}{\theta_i}\right) \tag{B.15}$$

$$= \inf_\theta \sum_{i=1}^d \left(\frac{|w_i|}{\theta_i}\right)^{p-1} |w_i| \frac{1}{p} \tag{B.16}$$

$$= \frac{1}{p} \inf_\theta \sum_{i=1}^d \frac{|w_i|^p}{\theta_i^{p-1}} \tag{B.17}$$

$$= \frac{1}{p}\|w\|^p, \tag{B.18}$$

where we have used $\frac{q}{q-1} = 1 + \frac{1}{q-1}$, $\frac{1}{q} = 1 - \frac{1}{p}$ and $\frac{1}{q-1} = p - 1$. It follows that the Fenchel conjugate of h^* is $\frac{1}{p}\|\cdot\|^p$, hence (B.12) and (B.13) are dual norms to each other as required. □

Appendix C

Appendix: Spectral Norms

In this appendix we present derivations for Chapter 4.

Proposition 71. *The unit ball of the spectral k-support norm is the convex hull of the set of matrices of rank at most k and Frobenius norm no greater than one.*

Proof. For any $W \in \mathbb{R}^{d\times T}$, define the following sets

$$T_k = \Big\{ W \in \mathbb{R}^{d\times T} \;\Big|\; \mathrm{rank}(W) \leq k,\ \|W\|_F \leq 1 \Big\}, \quad A_k = \mathrm{co}(T_k),$$

and consider the following functional

$$\lambda(W) = \inf\big\{ \lambda > 0 \;\big|\; W \in \lambda A_k \big\}, \quad W \in \mathbb{R}^{d\times T}. \tag{C.1}$$

We will apply Lemma 36 to the set A_k. To do this, we need to show that the set A_k is bounded, convex, symmetric and absorbing. The first three are clearly satisfied. To see that it is absorbing, let $W \in \mathbb{R}^{d\times T}$ have singular value decomposition $U\Sigma V^\top$, and let $r = \min(d, T)$. If W is zero then clearly $W \in A_k$, so assume it is non zero.

For $i \in \mathbb{N}_r$ let $S_i \in \mathbb{R}^{d\times T}$ have entry (i, i) equal to 1, and all remaining entries zero. We then have

$$W = U\Sigma V^\top = U\left(\sum_{i=1}^r \sigma_i S_i\right)V^\top = \left(\sum_{i=1}^d \sigma_i\right)\sum_{i=1}^r \frac{\sigma_i}{\sum_{j=1}^r \sigma_j}(US_iV^\top) =: \lambda\sum_{i=1}^r \beta_i Z_i.$$

Now for each i, $\|Z_i\|_F = \|S_i\|_F = 1$, and $\mathrm{rank}(Z_i) = \mathrm{rank}(S_i) = 1$, so $Z_i \in T_k$ for any $k \geq 1$. Furthermore $\beta_i \in [0, 1]$ and $\sum_{i=1}^r \beta_i = 1$, that is $(\beta_1, \ldots, \beta_r) \in \Delta^{r-1}$, so $\frac{1}{\lambda}W$ is a convex combination of Z_i, in other words $W \in \lambda A_k$, and we have shown that A_k is absorbing. It follows that A_k satisfies the hypotheses of Lemma 36, where we let $C = A_k$, hence λ defines a norm on $\mathbb{R}^{d\times T}$ with unit ball equal to A_k.

Since the constraints in T_k involve spectral functions, the sets T_k and A_k are invariant to left and right multiplication by orthogonal matrices. It follows that λ is a spectral function, that is $\lambda(W)$ is defined in terms of the singular values of W. By von Neumann's Theorem (Theorem 5) the norm it defines is orthogonally invariant and we have

$$\lambda(W) = \inf\left\{\lambda > 0 \,\middle|\, W \in \lambda A_k\right\} = \inf\left\{\lambda > 0 \,\middle|\, \sigma(W) \in \lambda C_k\right\} = \|\sigma(W)\|_{(k)}$$

where we have used Corollary 57, which states that C_k is the unit ball of the k-support norm. It follows that the norm defined by (C.1) is the spectral k-support norm with unit ball given by A_k. □

Lemma 73. *Let $r = \min(d,T)$ and let Θ be a bounded and convex subset of $\mathbb{R}^r_{++}$ which is invariant under permutation. For every $W = [w_1,\ldots,w_T] \in \mathbb{R}^{d\times T}$, it holds that*

$$\|W\Pi\|_\Theta = \min_{z\in\mathbb{R}^d} \|[w_1 - z,\ldots,w_T - z]\|_\Theta.$$

Proof. Given the set Θ we define the set $\Theta^{(T)} = \left\{\Sigma \in \mathbf{S}^T_{++} \,\middle|\, \lambda(\Sigma) \in \Theta\right\}$ and $\Theta^{(d)} = \left\{D \in \mathbf{S}^d_{++} \,\middle|\, \lambda(D) \in \Theta\right\}$. It follows from Lemma 8 that

$$\|W\|^2_\Theta \equiv \|\sigma(W)\|^2_\Theta = \inf_{\Sigma\in\Theta^{(T)}} \operatorname{tr}\left(\Sigma^{-1}W^\top W\right) = \inf_{D\in\Theta^{(d)}} \operatorname{tr}\left(D^{-1}WW^\top\right).$$

Using the second identity and recalling that $W\Pi = [w_1 - \bar{w},\ldots,w_T - \bar{w}]$, we have that

$$\begin{aligned}\|W\Pi\|^2_\Theta &= \inf_{D\in\Theta^{(d)}} \operatorname{tr}((W\Pi)^\top D^{-1}(W\Pi))\\ &= \inf_{D\in\Theta^{(d)}} \sum_{t=1}^T (w_t - \bar{w})^\top D^{-1}(w_t - \bar{w}) = \inf_{D\in\Theta^{(d)}} \min_{z\in\mathbb{R}^d} \sum_{t=1}^T (w_t - z)^\top D^{-1}(w_t - z)\end{aligned}$$

where in the last step we used the fact that the quadratic form $\sum_{t=1}^T (w_t - z)^\top D^{-1}(w_t - z)$ is minimized at $z = \bar{w}$. The result now follows by interchanging the infimum and the minimum in the last expression and using the definition of the Θ-norm. □

Appendix D

Appendix: The (k,p)-Support Norm

In this appendix we present derivations for Chapter 5.

Recall that the dual norm of a vector $u \in \mathbb{R}^d$ is defined by the optimization problem

$$\|u\|_{(k,p),*} = \max\big\{\langle u \mid w\rangle \;\big|\; \|w\|_{(k,p)} = 1\big\}. \tag{D.1}$$

Proposition 75. *If $p \in (1,\infty]$ then the dual (k,p)-support norm is given by*

$$\|u\|_{(k,p),*} = \left(\sum_{i\in I_k} |u_i|^q\right)^{\frac{1}{q}}, \quad u \in \mathbb{R}^d,$$

where $q = p/(p-1)$ and $I_k \subset \mathbb{N}_d$ is the set of indices of the k largest components of u in absolute value. Furthermore, if $p \in (1,\infty)$ and $u \in \mathbb{R}^d\backslash\{0\}$ then the maximum in (D.1) *is attained for*

$$w_i = \begin{cases} \operatorname{sign}(u_i)\left(\frac{|u_i|}{\|u\|_{(k,p),*}}\right)^{\frac{1}{p-1}} & \textit{if } i \in I_k, \\ 0 & \textit{otherwise}. \end{cases} \tag{D.2}$$

If $p = \infty$ the maximum is attained for

$$w_i = \begin{cases} \operatorname{sign}(u_i) & \textit{if } i \in I_k, u_i \neq 0, \\ \lambda_i \in [-1,1] & \textit{if } i \in I_k, u_i = 0, \\ 0 & \textit{otherwise}. \end{cases}$$

Proof. For every $u \in \mathbb{R}^d$ we have

$$\begin{aligned}
\|u\|_{(k,p),*} &= \max\left\{\sum_{i=1}^{d} u_i w_i \;\middle|\; w \in C_k^p\right\}\\
&= \max\left\{\sum_{i=1}^{d} u_i w_i \;\middle|\; \operatorname{card}(w) \le k, \|w\|_p \le 1\right\}\\
&= \max\left\{\sum_{i\in I_k} u_i w_i \;\middle|\; \sum_{i\in I_k} |w_i|^p \le 1\right\}\\
&= \left(\sum_{i\in I_k} |u_i|^q\right)^{\frac{1}{q}},
\end{aligned}$$

where the first equality uses the definition of the unit ball (5.5) and the second equality is true because the maximum of a linear functional on a compact set is attained at an extreme point of the set (Proposition 42). The third equality follows by using the cardinality constraint, that is we set $w_i = 0$ if $i \notin I_k$. Finally, the last equality follows by Hölder's inequality in $\mathbb{R}^k$, Proposition 3.

The second claim is a direct consequence of the cardinality constraint and Hölder's inequality in $\mathbb{R}^k$. □

Proposition 77. *The unit ball of the spectral (k,p)-support norm is the convex hull of the set of matrices of rank at most k and Schatten p-norm no greater than one.*

To prove Proposition 77 we follow the method of Proposition 71 and again refer to Lemma 36.

Proof. Define the set

$$T_k^p = \left\{W \in \mathbb{R}^{d\times m} \;\middle|\; \operatorname{rank}(W) \le k, \|\sigma(W)\|_p \le 1\right\},$$

and its convex hull $A_k^p = \operatorname{co}(T_k^p)$, and consider the Minkowski functional

$$\lambda(W) = \inf\left\{\lambda > 0 \;\middle|\; W \in \lambda A_k^p\right\}, \quad W \in \mathbb{R}^{d\times m}. \tag{D.3}$$

We show that A_k^p is absorbing, bounded, convex and symmetric, and it follows by Lemma 36 that λ defines a norm on $\mathbb{R}^{d\times m}$ with unit ball equal to A_k^p. The set A_k^p is clearly bounded, convex and symmetric. To see that it is absorbing, let W in $\mathbb{R}^{d\times m}$ have singular value decomposition $U\Sigma V^\top$, and let $r = \min(d,m)$. If W is zero then clearly $W \in A_k^p$, so assume it is non zero.

For $i \in \mathbb{N}_r$ let $S_i \in \mathbb{R}^{d\times m}$ have entry (i,i) equal to 1, and all remaining entries zero. We then have

$$\begin{aligned}
W &= U\Sigma V^\top \\
&= U\left(\sum_{i=1}^{r}\sigma_i S_i\right)V^\top \\
&= \left(\sum_{i=1}^{d}\sigma_i\right)\sum_{i=1}^{r}\frac{\sigma_i}{\sum_{j=1}^{r}\sigma_j}(US_iV^\top) \\
&=: \lambda\sum_{i=1}^{r}\lambda_i Z_i.
\end{aligned}$$

Now for each i, $\|\sigma(Z_i)\|_p = \|\sigma(S_i)\|_p = 1$, and $\text{rank}(Z_i) = \text{rank}(S_i) = 1$, so $Z_i \in T_k^p$ for any $k \geq 1$. Furthermore $\lambda_i \in [0,1]$ and $\sum_{i=1}^{r}\lambda_i = 1$, that is $(\lambda_1,\ldots,\lambda_r)$ lies in the unit simplex in $\mathbb{R}^d$, so $\frac{1}{\lambda}W$ is a convex combination of elements of Z_i. In other words $W \in \lambda A_k^p$, and we have shown that A_k^p is absorbing. It follows that A_k^p satisfies the hypotheses of Lemma 36, and λ defines a norm on $\mathbb{R}^{d\times m}$ with unit ball equal to A_k^p.

Since the constraints in T_k^p involve spectral functions, the sets T_k^p and A_k^p are invariant to left and right multiplication by orthogonal matrices. It follows that λ is a spectral function, that is $\lambda(W)$ is defined in terms of the singular values of W. By von Neumann's Theorem (Von Neumann, 1937) the norm it defines is orthogonally invariant and we have

$$\begin{aligned}
\lambda(W) &= \inf\left\{\lambda > 0 \;\middle|\; W \in \lambda A_k^p\right\} \\
&= \inf\left\{\lambda > 0 \;\middle|\; \sigma(W) \in \lambda C_k^p\right\} \\
&= \|\sigma(W)\|_{(k)}
\end{aligned}$$

where we have used Equation (5.5), which states that C_k^p is the unit ball of the (k,p)-support norm. It follows that the norm defined by (D.3) is the spectral (k,p)-support norm with unit ball given by A_k^p. □

Theorem 78. *Let $p \in (1,\infty)$. For every $w \in \mathbb{R}^d$, and $k \leq d$, it holds that*

$$\|w\|_{(k,p)} = \left[\sum_{i=1}^{\ell}(|w|_i^\downarrow)^p + \left(\frac{\sum_{i=\ell+1}^{d}|w|_i^\downarrow}{\sqrt[q]{k-\ell}}\right)^p\right]^{\frac{1}{p}} \qquad \text{(D.4)}$$

where $\frac{1}{p}+\frac{1}{q}=1$, and for $k=d$, we set $\ell=d$, otherwise ℓ is the largest integer in $\{0,\ldots,k-1\}$ satisfying

$$(k-\ell)|w|_\ell^\downarrow \geq \sum_{i=\ell+1}^{d}|w|_i^\downarrow. \qquad \text{(D.5)}$$

Furthermore, the norm can be computed in $\mathcal{O}(d\log d)$ time.

Proof. Note first that in (D.4) when $\ell = 0$ we understand the first term in the right hand side to be zero, and when $\ell = d+1$ we understand the second term to be zero.

We need to compute

$$\|w\|_{(k,p)} = \max\left\{\sum_{i=1}^{d} u_i w_i \;\middle|\; \|u\|_{(k,p),*} \leq 1\right\}$$

where the dual norm $\|\cdot\|_{(k,p),*}$ is described in Proposition 75. Let $z_i = |w|_i^{\downarrow}$. The problem is then equivalent to

$$\max\left\{\sum_{i=1}^{d} z_i u_i \;\middle|\; \sum_{i=1}^{k} u_i^q \leq 1, u_1 \geq \cdots \geq u_d\right\}. \tag{D.6}$$

This further simplifies to the k-dimensional problem

$$\max\left\{\sum_{i=1}^{k-1} u_i z_i + u_k \sum_{i=k}^{d} z_i \;\middle|\; \sum_{i=1}^{k} u_i^q \leq 1, u_1 \geq \cdots \geq u_k\right\}.$$

Note that when $k = d$, the solution is given by the dual of the ℓ_q-norm, that is the ℓ_p-norm. For the remainder of the proof we assume that $k < d$. We can now attempt to use Holder's inequality, which states that for all vectors x such that $\|x\|_q = 1$, $\langle x \mid y\rangle \leq \|y\|_p$, and the inequality is tight if and only if

$$x_i = \left(\frac{|y_i|}{\|y\|_p}\right)^{p-1} \operatorname{sign}(y_i).$$

We use it for the vector $y = (z_1, \ldots, z_{k-1}, \sum_{i=k}^{d} z_i)$. The components of the maximizer u satisfy $u_i = \left(\frac{z_i}{M_{k-1}}\right)^{p-1}$ if $i \leq k-1$, and

$$u_k = \left(\frac{\sum_{i=\ell+1}^{d} z_i}{M_{k-1}}\right)^{p-1}.$$

where for every $\ell \in \{0, \ldots, k-1\}$, M_ℓ denotes the r.h.s. in equation (D.4). We then need to verify that the ordering constraints are satisfied. This requires that

$$(z_{k-1})^{p-1} \geq \left(\sum_{i=k}^{d} z_i\right)^{p-1}$$

which is equivalent to inequality (D.5) for $\ell = k-1$. If this inequality is true we are done, otherwise we set $u_k = u_{k-1}$ and solve the smaller problem

$$\max\left\{\sum_{i=1}^{k-2} u_i z_i + u_{k-1}\sum_{i=k-1}^{d} z_i \;\middle|\; \sum_{i=1}^{k-2} u_i^q + 2u_{k-1}^q \le 1, \quad u_1 \ge \cdots \ge u_{k-1}\right\}.$$

We use again Hölder's inequality and keep the result if the ordering constraints are fulfilled. Continuing in this way, the generic problem we need to solve is

$$\max\left\{\sum_{i=1}^{\ell} u_i z_i + u_{\ell+1}\sum_{i=\ell+1}^{d} z_i \;\middle|\; \sum_{i=1}^{\ell} u_i^q + (k-\ell)u_{\ell+1}^q \le 1, \quad u_1 \ge \cdots \ge u_{\ell+1}\right\}$$

where $\ell \in \{0, \ldots, k-1\}$. Without the ordering constraints the maximum, M_ℓ, is obtained by the change of variable $u_{\ell+1} \mapsto (k-\ell)^{\frac{1}{q}} u_\ell$ followed by applying Hölder's inequality. A direct computation provides that the maximizer is $u_i = \left(\frac{z_i}{M_\ell}\right)^{p-1}$ if $i \le \ell$, and

$$(k-\ell)^{\frac{1}{q}} u_{\ell+1} = \left(\frac{\sum_{i=\ell+1}^{d} z_i}{(k-\ell)^{\frac{1}{q}} M_\ell^p}\right)^{p-1}.$$

Using the relationship $\frac{1}{p} + \frac{1}{q} = 1$, we can rewrite this as

$$u_{\ell+1} = \left(\frac{\sum_{i=\ell+1}^{d} z_i}{(k-\ell) M_\ell^p}\right)^{p-1}.$$

Hence, the ordering constraints are satisfied if

$$z_\ell^{p-1} \ge \left(\frac{\sum_{i=\ell+1}^{d} z_i}{k-\ell}\right)^{p-1},$$

which is equivalent to (D.5). Finally note that M_ℓ is a nondecreasing function of ℓ. This is because the problem with a smaller value of ℓ is more constrained, namely, it solves (D.6) with the additional constraints $u_{\ell+1} = \cdots = u_d$. Moreover, if the constraint (D.5) holds for some value $\ell \in \{0, \ldots, k-1\}$ then it also holds for a smaller value of ℓ, hence we maximize the objective by choosing the largest ℓ.

The computational complexity stems from using the monotonicity of M_ℓ with respect to ℓ, which allows us to identify the critical value of ℓ using binary search. □

Proposition 80. *For every $w \in \mathbb{R}^d$, the projection x of w onto the unit ball of the (k,∞)-norm is given by*

$$x_i = \begin{cases} \operatorname{sign}(w_i)(|w_i| - \beta), & \textit{if } ||w_i| - \beta| \le 1, \\ \operatorname{sign}(w_i), & \textit{if } ||w_i| - \beta| > 1, \end{cases} \tag{D.7}$$

where $\beta = 0$ if $\|w\|_1 \leq k$, otherwise $\beta \in (0,\infty)$ is chosen to maximize $\sum_{i=1}^d |x_i|$ subject to the constraint $\sum_{i=1}^d |x_i| \leq k$. Furthermore, the projection can be computed in $\mathcal{O}(d\log d)$ time.

Remark 80a. *The intuition behind the proof follows from the general case study, which comprises four cases for the two constraints, assuming all $w_i > 0$:*

1. $|w_i| \leq 1$ *for all i, and $\sum_i |w_i| < k$: set $x = w$, neither constraint will be tight;*
2. $|w_i| > 1$ *for some i, and $\sum_i |w_i| < k$: separable case, $x = sign(w)\min(|w|,1)$, ℓ_1 constraint will not be tight;*
3. $|w_i| \leq 1$ *for all i, and $\sum_i |w_i| \geq k$: line search for β, $x = sign(w)(|w_i| - \beta)$. ℓ_1 constraint will be tight;*
4. $|w_i| > 1$ *for some i, and $\sum_i |w_i| \geq k$: two subcases*
   (a) $\sum \min(|w_i|,1) < k$*: separable case after applying ℓ_∞ constraint, $x = sign(w)\min(|w_i|,1)$, ℓ_1 constraint will not be tight;*
   (b) $\sum \min(|w_i|,1) \geq k$*: line search for β, $x = sign(w)\min(|w_i| - \beta, 1)$, ℓ_1 constraint will be tight.*

Note that case 1 is a subset of case 2, and both can be grouped with case 4a. Similarly cases 3 and 4b can be grouped. Note that in 1, 2, and 4a the critical condition is $\sum \min(|w_i|,1) < k$, and in 3 and 4b it is $\sum \min(|w_i|,1) \geq k$. We can therefore write the above as

1. $\sum \min(|w_i|,1) < k$
   (a) $|w_i| \leq 1$ *for all i, $\sum_i |w_i| < k$: set $x = w$, neither constraint will be tight;*
   (b) $|w_i| > 1$ *some i, $\sum_i |w_i| < k$: separable case, $x = sign(w)\min(|w|,1)$, ℓ_1 constraint will not be tight;*
   (c) $|w_i| > 1$ *some i, $\sum_i |w_i| \geq k$: separable after applying ℓ_∞ constraint, $x = sign(w)\min(|w|,1)$, ℓ_1 constraint will not be tight.*
2. $\sum \min(|w_i|,1) \geq k$
   (a) $|w_i| \leq 1$ *for all i, $\sum_i |w_i| \geq k$: line search β, $x = sign(w)(|w_i| - \beta)$, ℓ_1 constraint will be tight;*
   (b) $|w_i| \leq 1$ *some i, $\sum_i |w_i| \geq k$: line search β, $x = sign(w)\min(|w_i| - \beta, 1)$, ℓ_1 constraint will be tight.*

Finally we note that in the last analysis, 1a, 1b, and 1c can be summarized in a single case, as can 2a and 2b.

1. *$\sum \min(|w_i|, 1) < k$: separable, if necessary after first applying ℓ_∞ constraint, $x =$ sign$(w) \min(|w_i|, 1)$, ℓ_1 constraint will not be tight;*

2. *$\sum \min(|w_i|, 1) \geq k$: line search β, $x =$ sign$(w) \min(|w_i| - \beta, 1)$, ℓ_1 constraint will be tight.*

The full proof of Proposition 80 follows.

Proof. We solve the optimization problem

$$\min_{x \in \mathbb{R}^d} \Big\{ \sum_{i=1}^{d} (x_i - w_i)^2 \;\Big|\; |x_i| \leq 1, \sum_{i=1}^{d} |x_i| \leq k \Big\} \tag{D.8}$$

by considering the different cases. When $\sum_{i=1}^{d} \min(|w_i|, 1) < k$, the solution is separable, and is given by

$$x_i = \begin{cases} -1, & \text{if } w_i < -1, \\ w_i, & \text{if } -1 \leq w_i \leq 1, \\ 1, & \text{if } w_i > 1, \end{cases} \tag{D.9}$$

or equivalently

$$x_i = \text{sign}(w_i) \min(|w_i|, 1). \tag{D.10}$$

Assuming $w_i \neq 0$ for all i, and $\sum_{i=1}^{d} \min(|w_i|, 1) \geq k$ the Lagrangian with nonnegative multiplier 2β becomes

$$\begin{aligned} \mathcal{L}(x, \beta) &= \frac{1}{2} \sum_{i=1}^{d} (x_i - w_i)^2 + 2\beta \left(\sum_{i=1}^{d} \min(|x_i|, 1) - k \right) \\ &= \frac{1}{2} \sum_{i=1}^{d} (x_i - w_i)^2 + \beta \left(\sum_{i=1}^{d} (|x_i| + 1 - ||x_i| - 1|) - k \right) \end{aligned}$$

using $\min(a, b) = \frac{1}{2}(a + b - |a - b|)$.

We find the derivative $\frac{\partial L}{\partial x_i} = x_i - w_i + \frac{1}{2}\beta \text{sign}(x_i)(1 - \text{sign}(|x_i| - 1))$. Setting to zero we get

$$x_i = w_i - \frac{1}{2}\beta \text{sign}(x_i)(1 - \text{sign}(|x_i| - 1)),$$

multiplying by sign(x_i) we get

$$|x_i| = \text{sign}(x_i) w_i - \frac{1}{2}\beta(1 - \text{sign}(|x_i| - 1)). \tag{D.11}$$

The second term on the RHS is nonnegative, so since the LHS is also nonnegative we need $\text{sign}(x_i) = \text{sign}(w_i)$. The first term on the RHS then becomes $\text{sign}(x_i)w_i = \text{sign}(w_i)w_i = |w_i|$, so we get

$$0 \leq |x_i| = |w_i| - \frac{1}{2}\beta(1 - \text{sign}(|x_i| - 1)), \tag{D.12}$$

hence we require

$$|w_i| \geq \frac{1}{2}\beta(1 - \text{sign}(|x_i| - 1)). \tag{D.13}$$

Considering the RHS of (D.13), note that when $|x_i| \leq 1$ the RHS is β, that is we require $|w_i| \geq \beta$. If $|x_i| > 1$, the RHS evaluates to zero, and (D.13) is trivially satisfied, however this case will be excluded by the ℓ_∞ constraint.

When $w_i \geq 0$, $x_i \geq 0$, and by assumption $w_i \geq \beta$, we have

$$x_i = w_i - \beta, \tag{D.14}$$

and imposing $|x_i| \leq 1$ this becomes

$$x_i = \begin{cases} w_i - \beta, & \text{if } 0 \leq w_i - \beta \leq 1, \\ 1, & \text{if } w_i - \beta > 1, \end{cases}. \tag{D.15}$$

The case $w_i \leq 0$ follows by symmetry and we find

$$x_i = \begin{cases} -1, & \text{if } w_i + \beta < -1, \\ w_i + \beta, & \text{if } -1 \leq w_i + \beta \leq 0, \\ w_i - \beta, & \text{if } 0 \leq w_i - \beta \leq 1, \\ 1, & \text{if } w_i - \beta > 1, \end{cases}, \tag{D.16}$$

hence we can concisely write (D.16) as

$$x_i = \text{sign}(w_i)(\min(|w_i| - \beta, 1). \tag{D.17}$$

where β is zero when $\sum_i \min(|w_i|, 1) < k$, otherwise it is chosen to maximize $\sum_i |x_i|$, specifically achieving $\sum_i \min(|x_i|, 1) = k$. □

Remark 80b. *The case where some components of w are zero follows from the case that we have considered above. Specifically, we form the vector w' from the non zero components of w and solve the corresponding problem using Proposition 80 to obtain x'. The solution x to the* *general problem is then given by identifying its components to those of* x' *on the support set of* w*, and setting the remaining components to zero, which maintains the optimality of the objective in* (D.8).

Appendix E

Appendix: The Tensor k-Support Norm

In this appendix we present derivations for Chapter 6.

Proposition 83. *There holds the bound*

$$\mathcal{R} \leq \frac{2}{m} \max_{1\leq n\leq N} k_n^{\frac{1}{q_n}} \mathbb{E}_\epsilon \Big\| M_n\Big(\sum_{j=1}^m \epsilon_j \boldsymbol{X}_j\Big)\Big\|_{\text{sp}} + 8C\sqrt{\frac{\log N}{m}}$$

where $q_n \in [1,\infty]$ *satisfies* $\frac{1}{p_n}+\frac{1}{q_n}=1$ *and*

$$C = \sqrt{\max_{n=1}^{N} \sup_{\|M_n(\boldsymbol{W})\|_{(k_n,p_n)}\leq 1} \frac{1}{m}\sum_{j=1}^m \langle \boldsymbol{W} \mid \boldsymbol{X}_j\rangle^2}.$$

In particular for tensor completion we have

$$\mathcal{R}_\Omega \leq 2k^*\left[\frac{1}{\sqrt{m}} + \sqrt{\frac{2\log(N+1)}{m}}\right] + 8\frac{\sqrt{\log N}}{m}$$

where $k^* = \max_{1\leq n\leq N} k_n^{\frac{1}{q_n}}$.

Proof. Recall the shorthand notation $\|\boldsymbol{W}\|_{[n]} = \|M_n(\boldsymbol{W})\|_{(k_n,p_n)}$ and $\|\boldsymbol{U}\|_{[n],*} = \|M_n(\boldsymbol{U})\|_{(k_n,p_n),*}$ introduced earlier. Using the definition of the dual norm we have

$$\begin{aligned}
\mathcal{R} &= \frac{2}{n}\mathbb{E}_\epsilon \sup_{|||\boldsymbol{W}|||\leq 1} \left\langle \sum_{j=1}^m \epsilon_j \mathbf{X}_j \mid \boldsymbol{W}\right\rangle \\
&= \frac{2}{n}\mathbb{E}_\epsilon \; |||\sum_{j=1}^m \epsilon_j \mathbf{X}_j|||_* \\
&= \frac{2}{n}\mathbb{E}_\epsilon \|\sum_{j=1}^m \epsilon_j \mathbf{X}_j\|_{[n],*} \\
&= \frac{2}{n}\max_{1\leq n\leq N} \mathbb{E}_\epsilon \|\sum_{j=1}^m \epsilon_j \mathbf{X}_j\|_{[n],*} + 8\sqrt{C\frac{\log N}{m}} \qquad \text{(E.1)}
\end{aligned}$$

where the last step follows by (Corollary 3 Maurer et al., 2014) with

$$C = \max_{1 \le n \le N} \sup_{\|\boldsymbol{W}\|_n \le 1} \frac{1}{n} \sum_{j=1}^{m} \langle \boldsymbol{W} \mid \mathbf{X}_j \rangle^2.$$

This proves the first bound.

In the tensor completion case a crude bound gives

$$\sum_{j=1}^{m} \langle \boldsymbol{W} \mid \mathbf{X}_j \rangle^2 = \sum_{(i_1,\ldots,i_N) \in \Omega} W^2_{i_1,\ldots,i_N} \le \|\boldsymbol{W}\|^2_{\text{Frob}} \le 1$$

bounding the second term in the r.h.s. of (E.1). To bound the first term, we let $\boldsymbol{D} = \sum_{j=1}^{m} \epsilon_j \mathbf{X}_j$ and note that

$$\mathbb{E}_\epsilon \| \sum_{j=1}^{m} \epsilon_j \mathbf{X}_j \|_{[n],*} \le \qquad \mathbb{E}_\epsilon \left\{ \sum_{\ell=1}^{k_n} \Big[\sigma_\ell\big(M_n(\boldsymbol{D})\big) \Big]^{q_n} \right\}^{\frac{1}{q_n}} \le k_n^{\frac{1}{q_n}} \mathbb{E}_\epsilon \|M_n(\boldsymbol{D})\|_{\text{sp}}.$$

It remains to bound $\mathbb{E}_\epsilon \|M_n(\boldsymbol{D})\|_{\text{sp}}$. To this end we use Proposition 6 of Maurer and Pontil (2013) which, adapted to our notation, gives

$$\mathbb{E}_\epsilon \|M_n(\boldsymbol{D})\|_{\text{sp}} \le \sqrt{\| \sum_t \sum_{j=1}^{m_t} x_j^t \otimes x_j^t \|_{\text{sp}}} + \sqrt{2 \max_t \left\{ m_t \max_{j=1}^{m_t} \|x_j^t\|^2 \right\} \log(N+1)}$$

where $t = (i_1, \ldots, i_{n-1}, i_{n+1}, \ldots, i_N)$ and $x_j^t \in \mathbb{R}^{d_n}$ is a vector the components of which are all equal to zero expect one component. This component, $i_n(j)$ is uniquely determined by the condition that $(i_1, \ldots, i_{n-1}, i_n(j), i_{n+1}, \ldots, i_N) \in \Omega$. Hence $\|x_j^t\|^2 = 1$ for every j and t. The sum over the index j runs over the set of observed tensor entries in which the multi-index t is fixed and the single index i_n varies. The result then follows by noting that $m_t \le m$ and

$$\| \sum_t \sum_{j=1}^{m_t} x_j^t \otimes x_j^t \|_{\text{sp}} \le \text{tr} \Big(\sum_t \sum_{j=1}^{m_t} x_j^t \otimes x_j^t \Big) = \sum_t \sum_{j=1}^{m_t} \|x_j^t\|^2 = m.$$

□

Appendix F

Appendix: Interpolation Norms

In this appendix we present derivations for Chapter 7. Recall first the following proposition and assumption.

Proposition 84. *Let m, d, and N be strictly positive integers. Let $K = \mathbb{R}^m_-$, let $B \in \mathbb{R}^{d\times N}$, and let $F\colon \mathbb{R}^N \to \mathbb{R}^m$ be K-convex in the sense that, for every $\alpha \in \left]0,1\right[$ and every $(u,v) \in \mathbb{R}^N \times \mathbb{R}^N$*

$$F\big(\alpha u + (1-\alpha)v\big) - \alpha F(u) - (1-\alpha)F(v) \in K. \tag{F.1}$$

Let $h\colon \mathbb{R}^m \to \left]-\infty,+\infty\right]$ be a proper convex function such that, for every $(x,y) \in \operatorname{ran} F \times \operatorname{ran} F$,

$$x - y \in K \quad \Rightarrow \quad h(x) \leq h(y). \tag{F.2}$$

Define, for every $w \in \mathbb{R}^d$,

$$\varphi(w) = \inf_{\substack{v\in\mathbb{R}^N\\ Bv=w}} h\big(F(v)\big). \tag{F.3}$$

Then $h\circ F$ and φ are convex.

Assumption 85. *Let m, d, and $(r_j)_{1\leq j\leq m}$ be strictly positive integers, set $N = \sum_{j=1}^m r_j$, and let $v = (v_j)_{1\leq j\leq m}$ denote a generic element in $\mathbb{R}^N$ where, for every $j \in \{1,\ldots,m\}$, $v_j \in \mathbb{R}^{r_j}$. Furthermore:*

1. *$F\colon v \mapsto (f_j(v_j))_{1\leq j\leq m}$ where, for every $j \in \{1,\ldots,m\}$, f_j is a norm on $\mathbb{R}^{r_j}$ with dual norm $f_{j,*}$.*
2. *$|||\cdot|||$ is a norm on $\mathbb{R}^m$ which is monotone in the sense that, for every $(x,y) \in \mathbb{R}^m_+ \times \mathbb{R}^m_+$,*

$$x - y \in \mathbb{R}^m_- \quad \Rightarrow \quad |||x||| \leq |||y|||, \tag{F.4}$$

   and $|||\cdot|||_*$ denotes its dual norm.

3. *For every $j \in \{1,\ldots,m\}$, $B_j \in \mathbb{R}^{d\times r_j}$, and $B = [B_1\ \cdots\ B_m]$ has full rank.*

Let, for every $w \in \mathbb{R}^d$,

$$\|w\| = \min_{\substack{v \in \mathbb{R}^N \\ Bv = w}} |||F(v)||| = \min_{\substack{v \in \mathbb{R}^N \\ \sum_{j=1}^m B_j v_j = w}} \big|\big|\big|\big(f_1(v_1), \dots, f_m(v_m)\big)\big|\big|\big|. \tag{F.5}$$

Proposition 86. *Consider the setting of Assumption 85. Then $|||\cdot||| \circ F$ is a norm on $\mathbb{R}^N$ and its dual norm at $t \in \mathbb{R}^N$ is*

$$(|||\cdot||| \circ F)_*(t) = \big|\big|\big|\big(f_{1,*}(t_1), \dots, f_{m,*}(t_m)\big)\big|\big|\big|. \tag{F.6}$$

Proof. Let v and t be in $\mathbb{R}^N$. We first deduce from Assumption 85.1 that

$$\begin{aligned}(\forall \alpha \in \mathbb{R}) \quad |||F(\alpha v)||| &= |||(f_1(\alpha v_1), \dots, f_m(\alpha v_m))||| \\ &= |||(|\alpha| f_1(v_1), \dots, |\alpha| f_m(v_m))||| \\ &= |\alpha| \, |||(f_1(v_1), \dots, f_m(v_m))||| \\ &= |\alpha| \, |||F(v)||| \end{aligned} \tag{F.7}$$

and that

$$\begin{aligned} |||F(v)||| = 0 \quad &\Leftrightarrow \quad F(v) = 0 \\ &\Leftrightarrow \quad (\forall j \in \{1, \dots, m\}) \;\; f_j(v_j) = 0 \\ &\Leftrightarrow \quad (\forall j \in \{1, \dots, m\}) \;\; v_j = 0 \\ &\Leftrightarrow \quad v = 0. \end{aligned} \tag{F.8}$$

Let us now check the triangle inequality. By Assumption 85.1, $F(v+t) - F(v) - F(t) \in \mathbb{R}^m_-$. Hence, we derive from (F.4) that

$$|||F(v+t)||| \leq |||F(v) + F(t)||| \leq |||F(v)||| + |||F(t)|||. \tag{F.9}$$

Finally, to establish (F.6), observe that

$$\langle v \mid t \rangle = \sum_{j=1}^m \langle v_j \mid t_j \rangle \leq \sum_{j=1}^m f_j(v_j) f_{j,*}(t_j) \leq |||F(v)||| \, |||F_*(t)|||_*. \tag{F.10}$$

Since for every t we can choose v so that both equalities are tight, the result follows. □

Recall further the folowing remark.

Remark 87. *Let $w \in \mathbb{R}^d$, set $C = \big\{v \in \mathbb{R}^N \;\big|\; Bv = w\big\}$, and let d_C be the distance function to C*

associated with the norm $|||\cdot|||\circ F$ *(see Proposition 86), that is,*

$$(\forall z\in\mathbb{R}^N)\quad d_C(z)=\inf_{v\in C}|||F(v-z)|||. \tag{F.11}$$

Note that C *is a closed affine subspace and it follows from* (F.11) *and* (F.5) *that*

$$d_C(0)=\inf_{v\in C}|||F(v-0)|||=\inf_{v\in C}|||F(v)|||=\|w\|. \tag{F.12}$$

Thus, the function $\|\cdot\|$ *in* (F.5) *is defined via a minimal norm interpolation process. Hence, the optimization problem underlying* (F.5) *is that of minimizing the norm* $|||\cdot|||\circ F$ *over the closed affine subspace* C*. It therefore possesses a solution, namely the minimum norm element in* C*.*

Theorem 88. *Consider the setting of Assumption 85. Then the following hold:*

1. $\|\cdot\|$ *is a norm.*

2. *The dual norm of* $\|\cdot\|$ *at* $u\in\mathbb{R}^d$ *is*

$$\|u\|_*=\Big|\Big|\Big|\big(f_{1,*}(B_1^\top u),\dots,f_{m,*}(B_m^\top u)\big)\Big|\Big|\Big|_*. \tag{F.13}$$

Proof. 1: We first note that, since $\operatorname{ran}B=\mathbb{R}^d$, $\operatorname{dom}\|\cdot\|=\mathbb{R}^d$. Next, we derive from (F.5) that, for every $w\in\mathbb{R}^d$ and every $\alpha\in\mathbb{R}$,

$$\|\alpha w\|=|\alpha|\,\|w\|. \tag{F.14}$$

On the other hand, it is clear that F satisfies (F.1), that $|||\cdot|||$ satisfies (F.2), and that (F.5) assumes the same form as (F.3). Hence, by Proposition 84 the function $\|\cdot\|$ is convex. In view of (F.14), we therefore have, for every $(w,t)\in\mathbb{R}^d\times\mathbb{R}^d$, $\|w+t\|\leq\|w\|+\|t\|$. Now let $w\in\mathbb{R}^d$ be such that $\|w\|=0$ and set $C=\{v\in\mathbb{R}^N \mid Bv=w\}$. Then it follows from (F.12) that $d_C(0)=0$ and, since C is closed, we get $0\in C$. Therefore, $w=B0=0$. Altogether $\|\cdot\|$ is a norm.

2: Let $u\in\mathbb{R}^d$. Then

$$\begin{aligned}
\|u\|_*&=\max\big\{\langle w\mid u\rangle \;\big|\; w\in\mathbb{R}^d,\ \|w\|\leq 1\big\}\\
&=\max\Big\{\langle w\mid u\rangle \;\Big|\; w\in\mathbb{R}^d,\ \min_{Bv=w}|||F(v)|||\leq 1\Big\}\\
&=\max\big\{\langle Bv\mid u\rangle \;\big|\; v\in\mathbb{R}^N,\ |||F(v)|||\leq 1\big\}\\
&=\max\big\{\langle v\mid B^\top u\rangle \;\big|\; v\in\mathbb{R}^N,\ |||F(v)|||\leq 1\big\}\\
&=\big(|||\cdot|||\circ F\big)_*(B^\top u),
\end{aligned} \tag{F.15}$$

where last step follows from Proposition 86. We conclude by applying (F.6). □

Derivation of the random block-coordinate algorithm

In this section we present the full derivation of the algorithm. Recall that we write the regularization problem as

$$\underset{\substack{w\in\mathbb{R}^d\\ v_1\in\mathbb{R}^{r_1},\ldots,v_m\in\mathbb{R}^{r_m}\\ \sum_{j=1}^m B_j v_j = w}}{\text{minimize}} \quad \sum_{i=1}^n \ell_i(\langle w \mid a_i\rangle, \beta_i) + \lambda \sum_{j=1}^m f_j(v_j). \tag{F.16}$$

We can therefore solve the optimization problem

$$\underset{v_1\in\mathbb{R}^{r_1},\ldots,v_m\in\mathbb{R}^{r_m}}{\text{minimize}} \quad \sum_{i=1}^n \ell_i\Bigg(\sum_{j=1}^m \langle B_j v_j \mid a_i\rangle, \beta_i\Bigg) + \lambda \sum_{j=1}^m f_j(v_j), \tag{F.17}$$

and derive from one of its solutions $(v_j)_{1\le j\le m}$ a solution $w = \sum_{j=1}^m B_j v_j$ to (F.16). To make the structure of this problem more apparent, let us introduce the functions

$$\Phi\colon (v_1,\ldots,v_m) \mapsto \lambda \sum_{j=1}^m f_j(v_j) \tag{F.18}$$

and

$$\Psi\colon (\eta_1,\ldots,\eta_n) \mapsto \sum_{i=1}^n \psi_i(\eta_i), \tag{F.19}$$

where, for every $i \in \{1,\ldots,n\}$,

$$\psi_i\colon \eta_i \mapsto \ell_i(\eta_i, \beta_i). \tag{F.20}$$

Let us also designate by $A \in \mathbb{R}^{n\times d}$ the matrix the rows of which are $(a_i)_{1\le i\le n}$ and define

$$L = [L_1 \cdots L_m] \in \mathbb{R}^{n\times N}, \tag{F.21}$$

where, for every $j \in \{1,\ldots,m\}$,

$$L_j = AB_j \in \mathbb{R}^{n\times r_j}. \tag{F.22}$$

Then $L = AB$, where $B = [B_1 \cdots B_m]$, and we can rewrite problem (F.17) as

$$\underset{v\in\mathbb{R}^N}{\text{minimize}} \;\; \Phi(v) + \Psi(Lv). \tag{F.23}$$

Note that in this concise formulation, the functions $\Phi \in \Gamma_0(\mathbb{R}^N)$ and $\Psi \in \Gamma_0(\mathbb{R}^n)$ are nonsmooth. Now let us introduce the functions

$$\boldsymbol{F}\colon (v,s) \mapsto \Phi(v) + \Psi(s) \quad \text{and} \quad \boldsymbol{G} = \iota_{\boldsymbol{V}}, \tag{F.24}$$

where $\boldsymbol{V} = \operatorname{gra} L = \big\{(v,s) \in \mathbb{R}^N \times \mathbb{R}^n \;\big|\; Lv = s\big\}$ is the graph of L, and $\iota_{\boldsymbol{V}}$ is the indicator function of $\boldsymbol{V}$. Using the variable $\boldsymbol{x} = (v,s)$, we see that (F.23) reduces to the problem

$$\underset{\boldsymbol{x}\in\mathbb{R}^{N+n}}{\text{minimize}}\; \boldsymbol{F}(\boldsymbol{x}) + \boldsymbol{G}(\boldsymbol{x}) \tag{F.25}$$

involving the sum of two functions in $\Gamma_0(\mathbb{R}^{N+n})$ and which can be solved with the Douglas-Rachford algorithm (Bauschke and Combettes, 2011, Section 27.2). Let $\boldsymbol{x}_0 \in \mathbb{R}^{N+n}$, let $\boldsymbol{y}_0 \in \mathbb{R}^{N+n}$, let $\boldsymbol{z}_0 \in \mathbb{R}^{N+n}$, let $\gamma \in \,]0,+\infty[$, and let $(\mu_k)_{k\in\mathbb{N}}$ be a sequence in $]0,2[$ such that $\inf_{k\in\mathbb{N}} \mu_n > 0$ and $\sup_{k\in\mathbb{N}} \mu_n < 2$. The Douglas-Rachford algorithm

```
for $k = 0, 1, \dots$
    $\boldsymbol{x}_{k+1} = \operatorname{prox}_{\gamma \boldsymbol{G}} \boldsymbol{y}_k$
    $\boldsymbol{z}_{k+1} = \operatorname{prox}_{\gamma \boldsymbol{F}} (2\boldsymbol{x}_{k+1} - \boldsymbol{y}_k)$
    $\boldsymbol{y}_{k+1} = \boldsymbol{y}_k + \mu_k(\boldsymbol{z}_{k+1} - \boldsymbol{x}_{k+1})$
```
(F.26)

produces a sequence $(\boldsymbol{x}_k)_{k\in\mathbb{N}}$ which converges to a solution to (F.25) (Bauschke and Combettes, 2011, Corollary 27.4). However,

$$\operatorname{prox}_{\boldsymbol{F}} \colon (v,s) \mapsto (\operatorname{prox}_{\Phi} v, \operatorname{prox}_{\Psi} s), \tag{F.27}$$

and

$$\operatorname{prox}_{\boldsymbol{G}} \colon (x,y) \mapsto (v, Lv), \quad \text{where } v = x - L^\top(\operatorname{Id} + LL^\top)^{-1}(Lx - y), \tag{F.28}$$

is the projection operator onto $\boldsymbol{V}$. Hence, upon setting $R = L^\top(\operatorname{Id} + LL^\top)^{-1}$, we can rewrite (F.26) as

```
for $k = 0, 1, \dots$
    $q_k = Lx_k - y_k$
    $v_{k+1} = x_k - Rq_k$
    $s_{k+1} = Lv_{k+1}$
    $z_{k+1} = \operatorname{prox}_{\gamma\Phi}(2v_{k+1} - x_k)$
    $t_{k+1} = \operatorname{prox}_{\gamma\Psi}(2s_{k+1} - y_k)$
    $x_{k+1} = x_k + \mu_k(z_{k+1} - v_{k+1})$
    $y_{k+1} = y_k + \mu_k(t_{k+1} - s_{k+1}),$
```
(F.29)

where we have set $\boldsymbol{x}_k = (v_k, s_k) \in \mathbb{R}^N \times \mathbb{R}^n$, $\boldsymbol{y}_k = (x_k, y_k) \in \mathbb{R}^N \times \mathbb{R}^n$, and $\boldsymbol{z}_k = (z_k, t_k) \in \mathbb{R}^N \times \mathbb{R}^n$. It follows from the above result that $(v_k)_{k\in\mathbb{N}}$ converges to a solution to (F.23). Let us now express (F.29) in terms of the original variables $(v_j)_{1\leq j\leq m}$ of (F.17). To this end set, for every $j \in \{1,\dots,m\}$,

$$R_j = L_j^\top(\mathrm{Id} + LL^\top)^{-1} = L_j^\top\left(\mathrm{Id} + \sum_{j=1}^{m} L_j L_j^\top\right)^{-1}. \tag{F.30}$$

Moreover, let us denote by $x_{j,k} \in \mathbb{R}^{r_j}$ the jth component of x_k, by $v_{j,k} \in \mathbb{R}^{r_j}$ the jth component of v_k, by $z_{j,k} \in \mathbb{R}^{r_j}$ the jth component of z_k. Furthermore, we denote by $\eta_{i,k} \in \mathbb{R}$ the ith component of y_k, by $\tau_{i,k} \in \mathbb{R}$ the ith component of t_k, and by $\sigma_{i,k} \in \mathbb{R}$ the ith component of s_k. Thus, (F.29) becomes

```
for $k = 0, 1, \dots$
    $q_k = \sum_{j=1}^{m} L_j x_{j,k} - y_k$
    for $j = 1, \dots, m$
        $v_{j,k+1} = x_{j,k} - R_j q_k$
        $z_{j,k+1} = \mathrm{prox}_{\gamma\lambda f_j}(2v_{j,k+1} - x_{j,k})$
        $x_{j,k+1} = x_{j,k} + \mu_k(z_{j,k+1} - v_{j,k+1})$                  (F.31)
    $s_{k+1} = \sum_{j=1}^{m} L_j v_{j,k+1}$
    for $i = 1, \dots, n$
        $\tau_{i,k+1} = \mathrm{prox}_{\gamma\psi_i}(2\sigma_{i,k+1} - \eta_{i,k})$
        $\eta_{i,k+1} = \eta_{i,k} + \mu_k(\tau_{i,k+1} - \sigma_{i,k+1})$.
```

In large-scale problems, a drawback of this approach is that $m+n$ proximity operators must be evaluated at each iteration, which can lead to impractical implementations in terms of computations and/or memory requirements. The analysis of (Combettes and Pesquet, 2015, Corollary 5.5) shows that the proximity operators in (F.31) can be sampled by sweeping through the indices in $\{1,\dots,m\}$ and $\{1,\dots,n\}$ randomly while preserving the convergence of the iterates. This results in partial updates of the variables which lead to significantly lighter iterations and remarkable flexibility in the implementation of the algorithm. Thus, a variable $v_{j,k}$ is updated at iteration k depending on whether a random activation variable $\varepsilon_{j,k}$ takes on the value 1 or 0 (each component $\eta_{i,k}$ of the vector y_k is randomly updated according to the same strategy). The method resulting from this random sampling scheme is presented in the next theorem.

Theorem 88. *Let* $\mathsf{D} = \{0,1\}^{m+n} \smallsetminus \{\mathsf{0}\}$*, let* $\gamma \in \left]0,+\infty\right[$*, let* $(\mu_k)_{k\in\mathbb{N}}$ *be a sequence in* $\left]0,2\right[$ *such that* $\inf_{k\in\mathbb{N}} \mu_n > 0$ *and* $\sup_{k\in\mathbb{N}} \mu_n < 2$*, let* $(v_{j,0})_{1\le j\le m}$ *and* $(x_{j,0})_{1\le j\le m}$ *be in* $\mathbb{R}^N$*, let* $y_0 = (\eta_{i,0})_{1\le i\le n} \in \mathbb{R}^n$*, and let* $(\boldsymbol{\varepsilon}_k)_{k\in\mathbb{N}} = (\varepsilon_{1,k}\dots,\varepsilon_{m+n,k})_{k\in\mathbb{N}}$ *be identically distributed* D-

valued random variables such that, for every $i \in \{1,\dots,m+n\}$, $\operatorname{Prob}[\varepsilon_{i,0}=1]>0$. *Iterate*

```
for $k = 0,1,\dots$
    $q_k = \sum_{j=1}^m L_j x_{j,k} - y_k$
    for $j = 1,\dots,m$
        $v_{j,k+1} = v_{j,k} + \varepsilon_{j,k}\big(x_{j,k} - R_j q_k - v_{j,k}\big)$
        $x_{j,k+1} = x_{j,k} + \varepsilon_{j,k}\mu_k\big(\operatorname{prox}_{\gamma\lambda f_j}(2v_{j,k+1} - x_{j,k}) - v_{j,k+1}\big)$        (F.32)
    $s_{k+1} = \sum_{j=1}^m L_j v_{j,k+1}$
    for $i = 1,\dots,n$
        $\eta_{i,k+1} = \eta_{i,k} + \varepsilon_{m+i,k}\mu_k\big(\operatorname{prox}_{\gamma\psi_i}(2\sigma_{i,k+1} - \eta_{i,k}) - \sigma_{i,k+1}\big).$
```

Suppose that the random vectors $(\boldsymbol{\varepsilon}_k)_{k\in\mathbb{N}}$ *and* $(x_k,y_k)_{k\in\mathbb{N}}$ *are independent. Then, for every* $j \in \{1,\dots,m\}$, $(v_{j,k})_{k\in\mathbb{N}}$ *converges almost surely to a vector* v_j *and* $w = \sum_{j=1}^m B_j v_j$ *is a solution to* (7.32).

Proof. It follows from (Combettes and Pesquet, 2015, Corollary 5.5) that $(v_{1,k},\dots,v_{m,k})_{k\in\mathbb{N}}$ converges almost surely to a solution to (F.17). In turn, w solves (F.16). □

Bibliography

Abernethy, J., Bach, F., Evgeniou, T., and Vert, J.-P. (2009). A new approach to collaborative filtering. *Journal of Machine Learning Research*, 10:803–826.

Aflalo, J., Ben-Tal, A., Bhattacharyya, C., Nath, J. S., and Raman, S. (2011). Variable sparsity kernel learninig. *Journal of Machine Learning Research*, 12:565–592.

Amelunxen, D., Lotz, M., McCoy, M. B., and Tropp, J. A. (2014). Living on the edge: A geometry theory of phase transitions in convex optimization. *Information and Inference*, 3(3):224–294.

Argyriou, A., Evgeniou, T., and Pontil, M. (2007a). Multi-task feature learning. *Advances in Neural Information Processing Systems 19*, pages 41–48.

Argyriou, A., Evgeniou, T., and Pontil, M. (2008). Convex multi-task feature learning. *Machine Learning*, 73(3):243–272.

Argyriou, A., Foygel, R., and Srebro, N. (2012). Sparse prediction with the k-support norm. *Advances in Neural Information Processing Systems 25*, pages 1466–1474.

Argyriou, A., Micchelli, C., Pontil, M., and Ying, Y. (2007b). A spectral regularization framework for multi-task structure learning. *Advances in Neural Information Processing Systems 20*, pages 25–32.

Argyriou, A., Micchelli, C. A., and Pontil, M. (2010). On spectral learning. *Journal of Machine Learning Research*, Vol. 11:905–923.

Argyriou, A., Micchelli, C. A., Pontil, M., Shen, L., and Xu, Y. (2011). Efficient first order methods for linear composite regularizers. *CoRR*, arXiv/1104.1436.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68.

Bach, F. (2010). Structured sparsity-inducing norms through submodular functions. In *Advances in Neural Information Processing Systems 23 (NIPS 2010)*, pages 118–126.

Bach, F. (2013). Learning with submodular functions: A convex optimization perspective. Technical report, INRIA - Ecole Normale Superieure.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2011). Convex optimization with sparsity-inducing norms. In Sra, S., Nowozin, S., and Wright, S. J., editors, *Optimization for Machine Learning, 1.*, chapter 4. The MIT Press.

Bach, F., Jenatton, R., Mairal, J., and Obozinski, G. (2012). Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106.

Bach, F., Mairal, J., and Ponce, J. (2008). Convex sparse matrix factorizations. *arXiv preprint, arXiv:0812.1869*.

Bartlett, P. L. and Mendelson, S. (2002). Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482.

Bauschke, H. H. and Combettes, P. L. (2011). *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer.

Baxter, J. (2000). A model of inductive bias learning. *Journal of Artificial Intelligence Research*, Vol. 12:149–198.

Beck, A. and Teboulle, M. (2009). A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202.

Becker, S. R. and Combettes, P. L. (2014). An algorithm for splitting parallel sums of linearly composed monotone operators, with applications to signal recovery. *Journal of Nonlinear and Convex Analysis*, 15(137–159).

Bertsekas, D. P. (1999). *Nonlinear Programming*. Athena Scientific.

Bertsekas, D. P. (2009). *Convex Optimization Theory*. Athena Scientific.

Bertsekas, D. P. (2015). *Convex Optimization Algorithms*. Athena Scientific.

Bertsekas, D. P., Nedic, A., and Ozdaglar, A. E. (2003). *Convex Analysis and Optimization*. Athena Scientific.

Bhatia, R. (1997). *Matrix Analysis*. Springer.

Bioucas-Dias, J. M. and Figueiredo, M. (2007). A new twist: Two-step iterative shrinkage/thresholding algorithms for image restoration. *IEEE Transactions on Image Processing*, 16(12):2992–3004.

Bogdan, J., Berg, E., Su, W., and Candes, E. (2013). Statistical estimation and testing via the ordered l1 norm. Technical Report 2013-07, Department of Statistics, Stanford Univerisity.

Borwein, J. M. and Lewis, A. S. (2000). *Convex Analysis and Nonlinear Optimization*. Springer-Verlag.

Borwein, J. M. and Vanderwerff, J. D. (2010). *Convex functions. Constructions, characterizations and counterexamples*. Cambridge University Press.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, 37(4).

Candes, E., Romberg, J., and Tao, T. (2006). Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):12207–1223.

Candes, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, 58(3):1–37.

Cavallanti, G., Cesa-Bianchi, N., and Gentile, C. (2010). Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 1:2901–2934.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. (2012a). Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967.

Chandrasekaran, V., Recht, B., Parrilo, P. A., and Willsky, A. S. (2012b). The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12-6.

Chatterjee, S., Chen, S., and Banerjee, A. (2014). Generalized Dantzig selector: application to the k-support norm. In *Advances in Neural Information Processing Systems 28*, pages 1934–1942.

Chen, S. and Banerjee, A. (2016). Structured matrix recovery via the generalized Dantzig selector (2016). Technical report, University of Minnesota.

Cheney, W. and Light, W. (2000). *A Course in Approximation Theory*. American Mathematical Society, Providence, RI.

Cichocki, A., Mandic, D., De Lathauwer, L., Zhou, G., Zhao, Q., Caiafa, C., and Phan, H. A. (2015). Tensor decompositions for signal processing applications: From two-way to multiway component analysis. *IEEE Signal Processing Magazine*, 32(2):145–163.

Combettes, P. L. (2004). Solving monotone inclusions via compositions of nonexpansive averaged operators. *Optimization*, 53(5-6).

Combettes, P. L. and Pesquet, J.-C. (2011). Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems*. Springer.

Combettes, P. L. and Pesquet, J.-C. (2015). Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM Journal on Optimization*, 25(2):1221–1248.

Combettes, P. L. and Wajs, V. R. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling & Simulation*, 4(4):1168–1200.

Dacorogna, B. and Maréchal, P. (2008). The role of perspective functions in convexity, polyconvexity, rank-one convexity and separate convexity. *Journal of Convex Analysis*, 15(2):271–284.

Do, C. B. and Ng, A. Y. (2005). Transfer learning for text classification. In *Advances in Neural Information Processing Systems 18*, pages 299–306.

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., and Darrell, T. (2014). DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In *Proceedings of the 31st International Conference on Machine Learning*, pages 647–655.

Donoho, D. (2006). Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306.

Drusvyatskiy, D., Vavasis, S. A., and Wolkowicz, H. (2015). Extreme point inequalities and geometry of the rank sparsity ball. *Mathematical Programming*, 152(1):521–544.

Evgeniou, T., Micchelli, C. A., and Pontil, M. (2005). Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637.

Evgeniou, T. and Pontil, M. (2004). Regularized multi–task learning. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 109–117.

Evgeniou, T., Pontil, M., and Toubia, O. (2007). A convex optimization approach to modeling heterogeneity in conjoint estimation. *Marketing Science*, 26:805–818.

Fazel, M. (2002). *Matrix Rank Minimization with Applications*. PhD thesis, Stanford University.

Fazel, M., Hindi, H., and Boyd, S. P. (2001). A rank minimization heuristic with application to minimum orders system approximation. *Proceedings of the American Control Conference*.

Figueiredo, M. A. T. and Nowak, R. D. (2016). Ordered weighted l1 regularized regression with strongly correlated covariates: theoretical aspects. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 930–938.

Fornasier, M. and Rauhut, H. (2008). Recovery algorithms for vector-valued data with joint sparsity constraints. *SIAM Journal on Numerical Analysis*, 46(2):577–613.

Frank, I. and Friedman, J. (1993). A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148.

Frank, M. and Wolfe, P. (1956). An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3 (1-2):95–110.

Friedland, S. and Lim, L.-H. (2016). Nuclear norm of higher-order tensors. *preprint available at arXiv: 1410.6072 v3*.

Gandy, S., Recht, B., and Yamada, I. (2011). Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2).

Gramfort, A. and Kowalski, M. (2009). Improving M/EEG source localization with an intercondition sparse prior. In *IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 141–144.

Grandvalet, Y. (1998). Least absolute shrinkage is equivalent to quadratic penalization. In *ICANN 98*, pages 201–206. Springer London.

Gunasekar, S., Banerjee, A., and Ghosh, J. (2015). Unified view of matrix completion under general structural constraints. In *Proceedings of Advances in Neural Information Processing Systems 28*, pages 1180–1188.

Hastie, T., Tibshirani, R., and Friedman, J. (2011). *The Elements of Statistical Learning*. Springer.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity*. CRC Press.

Hegde, C., Baraniuk, R., Davenport, M. A., and Duarte, M. F. (2014). An introduction to compressive sensing. OpenStax CNX. Aug 27, 2014 http://cnx.org/contents/f70b6ba0-b9f0-460f-8828-e8fc6179e65f@5.12.

Herbster, M. and Lever, G. (2009). Predicting the labelling of a graph via minimum p-seminorm interpolation. In *Proceedings of the 22nd Annual Conference on Learning Theory*.

Hillar, C. J. and Lim, L.-H. (2013). Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):45.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (1993). *Convex Analysis and Minimization Algorithms*. Springer-Verlag.

Hiriart-Urruty, J.-B. and Lemaréchal, C. (2001). *Fundamentals of Convex Analysis*. Springer.

Hoerl, A. and Kennard, R. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics,* Vol. 12, No. 1:pp. 55–67.

Horn, R. A. and Johnson, C. R. (1991). *Topics in Matrix Analysis*. Cambridge University Press.

Horn, R. A. and Johnson, C. R. (2005). *Matrix Analysis*. Cambridge University Press.

Huang, J., Zhang, T., and Metaxas, D. (2011). Learning with structured sparsity. *Journal of Machine Learning Research*, 12:3371–3412.

Ivanov, V., Vasin, V. V., and Tanana, V. (1978). *Theory of Linear Ill-Posed Problems and its Applications*. De Gruyter.

Jacob, L., Bach, F., and Vert, J.-P. (2009a). Clustered multi-task learning: a convex formulation. *Advances in Neural Information Processing Systems 21*, pages 745–752.

Jacob, L., Obozinski, G., and Vert, J.-P. (2009b). Group lasso with overlap and graph lasso. *Proceedings of the 26th International Conference on Machine Learning,* pages 433–440.

Jaggi, M. (2013). Revisiting Frank-Wolfe: Projection-free sparse convex optimization. *Proceedings of the 30th International Conference on Machine Learning*, pages 427–435.

Jaggi, M. and Sulovsky, M. (2010). A simple algorithm for nuclear norm regularized problems. *Proceedings of the 27th International Conference on Machine Learning*, pages 471–478.

Jalali, A., Fazel, M., and Xiao, L. (2016). Variational Gram functions: Convex analysis and optimization. *arXiv preprint, arXiv:1507.04734v2*.

Jenatton, R., Audibert, J., and Bach, F. (2011a). Structured variable selection with sparsity-inducing norms. *Journal of Machine Learning Research*, 12:2777–2824.

Jenatton, R., Mairal, J., Obozinski, G., and Bach, F. (2011b). Proximal methods for hierarchical sparse coding. *Journal of Machine Learning Research*, 12:2297–2334.

Jenatton, R., Obozinski, G., and Bach, F. (2010). Structured sparse principal component analysis. In *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 366–373.

Jojic, V., Saria, S., and Koller, D. (2011). Convex envelopes of complexity controlling penalties: the case against premature envelopment. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics*, Vol 15:399–406.

Kang, Z., Grauman, K., and Sha, F. (2011). Learning with whom to share in multi-task feature learning. In *Proceedings of the 28th International Conference on Machine Learning*, pages 521–528.

Karatzoglou, A., Amatriain, X., Baltrunas, L., and Oliver, N. (2010). Multiverse recommendation: n-dimensional tensor factorization for context-aware collaborative filtering. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 79–86. ACM.

Khan, J., Wei, J., Ringnér, M., Saal, L., Ladanyi, M., Westermann, F., Berthold, F., Schwab, C., Antonescu, C., Peterson, C., and Meltzer, P. (2001). Classification and diagnostic predition of cancers using gene expression profiling and aritificial neural networks. *Nature Medicine*, 7(6):673–9.

Kim, S. and Xing, E. P. (2010). Tree-guided group lasso for multi-task regression with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 543–550.

Kim, Y., Kim, J., and Kim, Y. (2006). Blockwise sparse regression. *Statistica Sinica*, 16:375–390.

Kloft, M., Ulf, B., Sonnenburg, S., and Zien, A. (2011). Lp-norm multiple kernel learning. *Journal of Machine Learning Research*, 12(953–997).

Kolda, T. G. and Bader, B. W. (2009). Tensor decompositions and applications. *SIAM Review*, 51(3):455–500.

Koltchinskii, V. and Panchenko, D. (2002). Empirical margin distributions and bounding the generalization error of combined classifiers. *The Annals of Statistics*, 30(1):1–50.

Körner, M.-C. (2011). *Minisum Hyperspheres*. Springer.

Kowalski, M. (2009). Sparse regression using mixed norms. *Applied and Computational Harmonic Analysis*, 27(3):303–324.

Krishnapuram, B., Carin, L., Figueiredo, M. A., and Hartemink, A. J. (2005). Sparse multinomial logistic regression: fast algorithms and generalization bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(6):957–68.

Lampert, C. H., Nickisch, H., and Harmeling, S. (2009). Learning to detect unseen object classes by between-class attribute transfer. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 951–958.

Lenk, P. J., DeSarbo, W. S., Green, P. E., and Young, M. R. (1996). Hierarchical Bayes conjoint analysis: Recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173–191.

Lewis, A. S. (1995). The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2:173–183.

Lewis, A. S. (2003). The mathematics of eigenvalue optimization. *Mathematical Programming*, 97(1-2):155–176.

Li, H., Chen, N., and Li, L. (2012). Error analysis for matrix elastic-net regularization algorithms. *IEEE Transactions on Neural Networks and Learning Systems*, 23-5:737–748.

Liu, J., Musialski, P., Wonka, P., and Ye, J. (2009). Tensor completion for estimating missing values in visual data. In *Proc. 12th International Conference on Computer Vision*, pages 2114–2121.

Louditski, A. (2016). Convex Optimization I: Introduction. http://ljk.imag.fr/membres/Anatoli.Iouditski/optimisation-convexe.htm.

Maréchal, P. (2005). On a functional operation generating convex functions, part 1: Duality. *Journal of Optimization Theory and Applications*, 126(1):175–189.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and its Applications*. Academic Press.

Maurer, A. and Pontil, M. (2012). Structured sparsity and generalization. *The Journal of Machine Learning Research*, 13:671–690.

Maurer, A. and Pontil, M. (2013). Excess risk bounds for multitask learning with trace norm regularization. In *Proceedings of The 27th Conference on Learning Theory (COLT)*.

Maurer, A., Pontil, M., and Romera-Paredes, B. (2014). An inequality with applications to structured sparsity and multitask dictionary learning. In *Proceedings of The 27th Conference on Learning Theory (COLT)*.

Mazumder, R., Hastie, T., and Tibshirani, R. (2010). Spectral regularization algorithms for learning large incomplete matrices. *Journal of Machine Learning Research*, 11:2287–2322.

McDonald, A., Pontil, M., and Stamos, D. (2016a). Fitting spectral decay with the k-support norm. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS) 2016*, pages 1061–1069.

McDonald, A. M., Pontil, M., and Stamos, D. (2014a). New perspectives on k-support and cluster norms. Technical report, University College London.

McDonald, A. M., Pontil, M., and Stamos, D. (2014b). Spectral k-support regularization. In *Advances in Neural Information Processing Systems 28*, pages 3644–3652.

McDonald, A. M., Pontil, M., and Stamos, D. (2016b). New perspectives on k-support and cluster norms. *Journal of Machine Learning Research*, 17(155):1–38.

Micchelli, C. A., Morales, J. M., and Pontil, M. (2010). A family of penalty functions for structured sparsity. *Advances in Neural Information Processing Systems 23*, pages 1612–1623.

Micchelli, C. A., Morales, J. M., and Pontil, M. (2013). Regularizers for structured sparsity. *Advances in Comp. Mathematics*, 38:455–489.

Micchelli, C. A. and Pontil, M. (2005). Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125.

Micchelli, C. A. and Pontil, M. (2007). Feature space perspectives for learning the kernel. *Machine Learning*, 66:297–319.

Micchelli, C. A., Shen, L., and Xu, Y. (2011). Proximity algorithms for image models: denoising. *Inverse Problems*, 27(045009).

Mirsky, L. (1960). Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, Vol. 11, Issue 1:50–59.

Moreau, J. (1965). Proximité et dualtité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299.

Mu, C., Huang, B., Wright, J., and Goldfarb, D. (2014). Square deal: Lower bounds and improved relaxations for tensor recovery. In *ICML*, pages 73–81.

Natarajan, B. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24:227–234.

Negahban, S. N. and Wainwright, M. J. (2011). Simultaneous support recovery in high dimensions: benefits and perils of block l1/linf regularization. *IEEE Transactions on Information Theory*, 57(6):3841–3863.

Negrinho, R. and Martins, A. F. T. (2014). Orbit regularization. In *Advances in Neural Information Processing Systems 27 (NIPS 2014)*, pages 3221–3229.

Nesterov, Y. (2007). Gradient methods for minimizing composite objective function. *Center for Operations Research and Econometrics*, 76.

Obozinski, G. and Bach, F. (2012). Convex relaxation for combinatorial penalties. *CoRR*, arXiv/1205.1240.

Obozinski, G. and Bach, F. (2016). A unified perspective on convex structured sparsity: Hierarchical, symmetric, submodular norms and beyond. Technical report, Université Paris-Est, Ecole Normale Supérieure. <hal-01412385>.

Obozinski, G., Taskar, B., and Jordan, M. (2010). Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 20(2):231–252.

Oymak, S., Jalali, A., Fazel, M., Eldar, Y. C., and Hassibi, B. (2015). Simultaneously structured models with application to sparse and low-rank matrices. *IEEE Transactions on Information Theory*, 61(5).

Pan, S. J. and Yang, Q. (2010). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.*, 22(10):1345–1359.

Parikh, N. and Boyd, S. (2013). Proximal algorithms. *Foundations and Trends in Optimization*, Vol. 1, No. 3:pp. 1–108.

Qaisar, S., Bilal, R. M., Iqbal, W., Naureen, M., and Lee, S. (2013). Compressive sensing: From theory to applications, a survey. *Journal of Communications and Networks*, 15(5):443–456.

Rapaport, F., Barillot, E., and Vert, J.-P. (2008). Classification of arrayCGH data using fused svm. *Bioinformatics*, 24(13):375–382.

Richard, E., Obozinksi, G., and Vert, J.-P. (2014). Tight convex relaxations for sparse matrix factorization. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3284–3292.

Rockafellar, R. T. (1970). *Convex Analysis*. Princeton University Press.

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational Analysis*. Springer.

Romera-Paredes (2014). *Multitask and transfer learning for multi-aspect data*. PhD thesis, University College London.

Romera-Paredes, B., Aung, H., Bianchi-Berthouze, N., and Pontil, M. (2013). Multilinear multitask learning. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1444–1452.

Romera-Paredes, B. and Pontil, M. (2013). A new convex relaxation for tensor completion. In *Advances in Neural Information Processing Systems 26 (NIPS 2013)*, pages 2967–2975.

Rudin, W. (1991). *Functional Analysis*. McGraw Hill.

Savalle, P.-A. (2014). *Interactions between rank and sparsity in penalized estimation, and detection of structured objects*. PhD thesis, Ecole Centrale Paris, https://tel.archives-ouvertes.fr/tel-01127356.

Signoretto, M., Dinh, Q. T., De Lathauwer, L., and Suykens, J. A. (2014). Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning*, 94(3):303–351.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2).

Srebro, N., Rennie, J. D. M., and Jaakkola, T. S. (2005). Maximum-margin matrix factorization. In *Advances in Neural Information Processing Systems 17*, pages 1329–1336.

Sun, S. (2008). Multitask learning for EEG-based biometrics. In *19th International Conference on Pattern Recognition*, pages 1–4.

Szafranski, M., Grandvalet, Y., and Morizet-Mahoudeaux, P. (2007). Hierarchical penalization. In *Advances in Neural Information Processing Systems 21*, pages 1457–1464.

Teschke, G. and Ramlau, R. (2007). An iterative algorithm for nonlinear inverse problems with joint sparsity constraints in vector-valued regimes and an application to color image inpainting. *Inverse Problems*, 23.

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

Tibshirani, R. and Saunders, M. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67:91–108.

Tikhonov, A. and Arsenin, V. (1977). *Solutions of Ill-Posed Problems*. Winston.

Toh, K.-C. and Yun, S. (2011). An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *SIAM J. on Img. Sci.*, 4:573–596.

Tomioka, R., Sukuki, T., Hayashi, K., and Kashima, H. (2011). Statistical performance of convex tensor decomposition. In *NIPS,* pages 972–980.

Tomioka, R. and Suzuki, T. (2013). Convex tensor decomposition via structured Schatten norm regularization. In *NIPS*, pages 1331–1339.

Torralba, A., Murphy, K. P., and Freeman, W. T. (2004). Sharing features: Efficient boosting procedures for multiclass object detection. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 762–769.

Tropp, J. A., Gilbert, A., and Strauss, M. (2006). Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(special issue "Sparse approximations in signal and image processing"):572–588.

Turlach, B., Venables, W., and Wright, S. (2005). Simultaneous variable selection. *Technometrics*, 47(3):349–363.

Vasin, V. V. (1970). Relationship of several variational methods for the approximate solution of ill-posed problems. *Matematicheskie Zametki*, Vol. 7, No. 3:265–272.

Vidal, G. (2003). Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14):147902.

Villa, S. (2014). Teaching notes: Proximal methods for machine learning. http://lcsl.mit.edu/data/silviavilla/Teaching.html.

Von Neumann, J. (1937). *Some matrix-inequalities and metrization of matric-space*. Tomsk. Univ. Rev. Vol I.

Wainwright, M. (2014). Structured regularizers for high-dimensional problems. *Annual Review of Statistics and Its Application*, 1:233–253.

Wimalawarne, K., Sugiama, M., and Tomioka, R. (2014). Multitask learning meets tensor factorization: task imputation via convex optimization. In *Advances in Neural Information Processing Systems 27 (NIPS)*.

Wright, S. J., Nowak, R. D., and Figueiredo, M. A. T. (2009). Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 52(7).

Yuan, M., Joseph, V. R., and Zou, H. (2009). Structured variable selection and estimation. *The Annals of Applied Statistics*, 3(4):1738–1757.

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

Zeng, X. and Figueiredo, M. A. T. (2014). Decreasing weighted sorted l1 regularization. *IEEE Signal Processing Letters*, 21:1240–1244.

Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, Vol. 37 and No. 6A.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320.