Optimization with Sparsity-Inducing Penalties

Optimization with Sparsity-Inducing Penalties

Francis Bach
INRIA — SIERRA Project-Team
France
francis.bach@inria.fr

Rodolphe Jenatton
INRIA — SIERRA Project-Team
France
rodolphe.jenatton@inria.fr

Julien Mairal
University of California, Berkeley
USA
julien@stat.berkeley.edu

Guillaume Obozinski
INRIA — SIERRA Project-Team
France
guillaume.obozinski@inria.fr

now
the essence of knowledge
Boston – Delft

Foundations and Trends® in Machine Learning

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
USA
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is F. Bach, R. Jenatton, J. Mairal and G. Obozinski, Optimization with Sparsity-Inducing Penalties, Foundation and Trends® in Machine Learning, vol 4, no 1, pp 1–106, 2011.

ISBN: 978-1-60198-510-1

Foundations and Trends® in Machine Learning
Volume 4 Issue 1, 2011

Editorial Board

Editorial Scope

Foundations and Trends® in Machine Learning will publish survey and tutorial articles in the following topics:

- Adaptive control and signal processing
- Applications and case studies
- Behavioral, cognitive and neural learning
- Bayesian learning
- Classification and prediction
- Clustering
- Data mining
- Dimensionality reduction
- Evaluation
- Game theoretic learning
- Graphical models
- Independent component analysis
- Inductive logic programming
- Kernel methods
- Markov chain Monte Carlo
- Model choice
- Nonparametric methods
- Online learning
- Optimization
- Reinforcement learning
- Relational learning
- Robustness
- Spectral methods
- Statistical learning theory
- Variational inference
- Visualization

Information for Librarians
Foundations and Trends® in Machine Learning, 2011, Volume 4, 4 issues. ISSN paper version 1935-8237. ISSN online version 1935-8245. Also available as a combined paper and online subscription.

Foundations and Trends® in
Machine Learning
Vol. 4, No. 1 (2011) 1–106

DOI: 10.1561/2200000015

now
the essence of knowledge

Optimization with Sparsity-Inducing Penalties

Francis Bach[1], Rodolphe Jenatton[2], Julien Mairal[3] and Guillaume Obozinski[4]

[1] *INRIA — SIERRA Project-Team, Laboratoire d'Informatique de l'Ecole Normale Supérieure, 23, avenue d'Italie, Paris, 75013, France, francis.bach@inria.fr*
[2] *INRIA — SIERRA Project-Team, rodolphe.jenatton@inria.fr*
[3] *Department of Statistics, University of California, Berkeley, CA 94720-1776, USA, julien@stat.berkeley.edu*
[4] *INRIA — SIERRA Project-Team, guillaume.obozinski@inria.fr*

Abstract

Sparse estimation methods are aimed at using or obtaining parsimonious representations of data or models. They were first dedicated to linear variable selection but numerous extensions have now emerged such as structured sparsity or kernel selection. It turns out that many of the related estimation problems can be cast as convex optimization problems by regularizing the empirical risk with appropriate nonsmooth norms. The goal of this monograph is to present from a general perspective optimization tools and techniques dedicated to such sparsity-inducing penalties. We cover proximal methods, block-coordinate descent, reweighted ℓ_2-penalized techniques, working-set and homotopy methods, as well as non-convex formulations and extensions, and provide an extensive set of experiments to compare various algorithms from a computational point of view.

Contents

1

Introduction

The principle of parsimony is central to many areas of science: the simplest explanation of a given phenomenon should be preferred over more complicated ones. In the context of machine learning, it takes the form of variable or feature selection, and it is commonly used in two situations. First, to make the model or the prediction more interpretable or computationally cheaper to use, i.e., even if the underlying problem is not sparse, one looks for the best sparse approximation. Second, sparsity can also be used given prior knowledge that the model should be sparse.

For variable selection in linear models, parsimony may be directly achieved by penalization of the empirical risk or the log-likelihood by the cardinality of the support[1] of the weight vector. However, this leads to hard combinatorial problems (see, e.g., [96, 136]). A traditional convex approximation of the problem is to replace the cardinality of the support by the ℓ_1-norm. Estimators may then be obtained as solutions of convex programs.

Casting sparse estimation as convex optimization problems has two main benefits: First, it leads to efficient estimation algorithms — and

[1] We call the set of non-zeros entries of a vector the support.

this monograph focuses primarily on these. Second, it allows a fruitful theoretical analysis answering fundamental questions related to estimation consistency, prediction efficiency [19, 99] or model consistency [145, 158]. In particular, when the sparse model is assumed to be well-specified, regularization by the ℓ_1-norm is adapted to high-dimensional problems, where the number of variables to learn from may be exponential in the number of observations.

Reducing parsimony to finding the model of lowest cardinality turns out to be limiting, and *structured parsimony* [15, 62, 64, 66] has emerged as a natural extension, with applications to computer vision [32, 62, 70], text processing [68], bioinformatics [64, 73] or audio processing [80]. Structured sparsity may be achieved by penalizing other functions than the cardinality of the support or regularizing by other norms than the ℓ_1-norm. In this monograph, we focus not only on norms which can be written as linear combinations of norms on subsets of variables, but we also consider traditional extensions such as multiple kernel learning and spectral norms on matrices (see Sections 1.3 and 1.5). One main objective of this monograph is to present methods which are adapted to most sparsity-inducing norms with loss functions potentially beyond least-squares.

Finally, similar tools are used in other communities such as signal processing. While the objectives and the problem set-ups are different, the resulting convex optimization problems are often similar, and most of the techniques reviewed in this monograph also apply to sparse estimation problems in signal processing. Moreover, we consider in Section 7 non-convex formulations and extensions.

This monograph aims at providing a general overview of the main optimization techniques that have emerged as most relevant and efficient for methods of variable selection based on sparsity-inducing norms. We survey and compare several algorithmic approaches as they apply not only to the ℓ_1-norm, group norms, but also to norms inducing structured sparsity and to general multiple kernel learning problems. We complement these by a presentation of some greedy and nonconvex methods. Our presentation is essentially based on existing literature, but the process of constructing a general framework leads naturally to new results, connections and points of view.

This monograph is organized as follows:

Sections 1.1 and 1.2 introduce respectively the notations used throughout the monograph and the optimization problem (1.1) which is central to the learning framework that we will consider.

Section 1.3 gives an overview of common sparsity and structured sparsity-inducing norms, with some of their properties and examples of structures which they can encode.

Section 1.4 provides an essentially self-contained presentation of concepts and tools from convex analysis that will be needed in the rest of the monograph, and which are relevant to understand algorithms for solving the main optimization problem (1.1). Specifically, since sparsity-inducing norms are nondifferentiable convex functions,[2] we introduce relevant elements of subgradient theory and Fenchel duality — which are particularly well suited to formulate the optimality conditions associated to learning problems regularized with these norms. We also introduce a general quadratic variational formulation for a certain class of norms in Section 1.4.2; the part on subquadratic norms is essentially relevant in view of sections on structured multiple kernel learning and can safely be skipped in a first reading.

Section 1.5 introduces *multiple kernel learning* (MKL) and shows that it can be interpreted as an extension of plain sparsity to reproducing kernel Hilbert spaces (RKHS), but formulated in the dual. This connection is further exploited in Section 1.5.2, where it is shown how structured counterparts of MKL can be associated with structured sparsity-inducing norms. These sections rely on Section 1.4.2. All sections on MKL can be skipped in a first reading.

In Section 2, we discuss classical approaches to solving the optimization problem arising from simple sparsity-inducing norms, such as interior point methods and subgradient descent, and point at their shortcomings in the context of machine learning.

Section 3 is devoted to a simple presentation of proximal methods. After two short sections introducing the main concepts and algorithms, the longer Section 3.3 focusses on the *proximal operator* and presents

[2] Throughout this monograph, we refer to sparsity-inducing norms such as the ℓ_1-norm as nonsmooth norms; note that all norms are nondifferentiable at zero, but some norms have more nondifferentiability points (see more details in Section 1.3).

algorithms to compute it for a variety of norms. Section 3.4 shows how proximal methods for structured norms extend naturally to the RKHS/MKL setting.

Section 4 presents block coordinate descent algorithms, which provide an efficient alternative to proximal method for *separable* norms like the ℓ_1- and ℓ_1/ℓ_2-norms, and can be applied to MKL. This section uses the concept of proximal operator introduced in Section 3.

Section 5 presents reweighted-ℓ_2 algorithms that are based on the quadratic variational formulations introduced in Section 1.4.2. These algorithms are particularly relevant for the least-squares loss, for which they take the form of iterative reweighted least-squares algorithms (IRLS). Section 5.2 presents a generally applicable quadratic variational formulation for general norms that extends the variational formulation of Section 1.4.2.

Section 6 covers algorithmic schemes that take advantage computationally of the sparsity of the solution by extending the support of the solution gradually. These schemes are particularly relevant to construct approximate or exact regularization paths of solutions for a range of values of the regularization parameter. Specifically, Section 6.1 presents working-set techniques, which are meta-algorithms that can be used with the optimization schemes presented in all the previous sections. Section 6.2 focuses on the homotopy algorithm, which can efficiently construct the entire regularization path of the Lasso.

Section 7 presents nonconvex as well as Bayesian approaches that provide alternatives to, or extensions of the convex methods that were presented in the previous sections. More precisely, Section 7.1 presents so-called greedy algorithms, that aim at solving the cardinality-constrained problem and include matching pursuit, orthogonal matching pursuit and forward selection; Section 7.2 presents continuous optimization problems, in which the penalty is chosen to be closer to the so-called ℓ_0-penalty (i.e., a penalization of the cardinality of the model regardless of the amplitude of the coefficients) at the expense of losing convexity, and corresponding optimization schemes. Section 7.3 discusses the application of sparse norms regularization to the problem of matrix factorization, which is intrinsically nonconvex, but for which the algorithms presented in the rest of this monograph are relevant.

Finally, we discuss briefly in Section 7.4 Bayesian approaches to sparsity and the relations to sparsity-inducing norms.

Section 8 presents experiments comparing the performance of the algorithms presented in Sections 2, 3, 4, 5, in terms of speed of convergence of the algorithms. Precisely, Section 8.1 is devoted to the ℓ_1-regularization case, and Sections 8.2 and 8.3 are respectively covering the ℓ_1/ℓ_p-norms with disjoint groups and to more general structured cases.

We discuss briefly methods and cases which were not covered in the rest of the monograph in Section 9 and we conclude in Section 10.

Some of the material from this monograph is taken from an earlier book chapter [12] and the dissertations of Rodolphe Jenatton [65] and Julien Mairal [85].

1.1 Notation

Vectors are denoted by bold lower case letters and matrices by upper case ones. We define for $q \geq 1$ the ℓ_q-norm of a vector $\boldsymbol{x}$ in $\mathbb{R}^n$ as $\|\boldsymbol{x}\|_q := (\sum_{i=1}^n |\boldsymbol{x}_i|^q)^{1/q}$, where $\boldsymbol{x}_i$ denotes the ith coordinate of $\boldsymbol{x}$, and $\|\boldsymbol{x}\|_\infty := \max_{i=1,\ldots,n} |\boldsymbol{x}_i| = \lim_{q\to\infty} \|\boldsymbol{x}\|_q$. We also define the ℓ_0-penalty as the number of nonzero elements in a vector[3]: $\|\boldsymbol{x}\|_0 := \#\{i \text{ s.t. } \boldsymbol{x}_i \neq 0\} = \lim_{q\to 0^+} (\sum_{i=1}^n |\boldsymbol{x}_i|^q)$. We consider the Frobenius norm of a matrix $\boldsymbol{X}$ in $\mathbb{R}^{m\times n}$: $\|\boldsymbol{X}\|_{\mathrm{F}} := (\sum_{i=1}^m \sum_{j=1}^n \boldsymbol{X}_{ij}^2)^{1/2}$, where $\boldsymbol{X}_{ij}$ denotes the entry of $\boldsymbol{X}$ at row i and column j. For an integer $n > 0$, and for any subset $J \subseteq \{1,\ldots,n\}$, we denote by $\boldsymbol{x}_J$ the vector of size $|J|$ containing the entries of a vector $\boldsymbol{x}$ in $\mathbb{R}^n$ indexed by J, and by $\boldsymbol{X}_J$ the matrix in $\mathbb{R}^{m\times|J|}$ containing the $|J|$ columns of a matrix $\boldsymbol{X}$ in $\mathbb{R}^{m\times n}$ indexed by J.

1.2 Loss Functions

We consider in this monograph convex optimization problems of the form

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w}), \tag{1.1}$$

[3] Note that it would be more proper to write $\|\boldsymbol{x}\|_0^0$ instead of $\|\boldsymbol{x}\|_0$ to be consistent with the traditional notation $\|\boldsymbol{x}\|_q$. However, for the sake of simplicity, we will keep this notation unchanged in the rest of the monograph.

where $f: \mathbb{R}^p \to \mathbb{R}$ is a convex differentiable function and $\Omega: \mathbb{R}^p \to \mathbb{R}$ is a sparsity-inducing — typically nonsmooth and non-Euclidean — norm.

In supervised learning, we predict outputs y in $\mathcal{Y}$ from observations $\boldsymbol{x}$ in $\mathcal{X}$; these observations are usually represented by p-dimensional vectors with $\mathcal{X} = \mathbb{R}^p$. In this supervised setting, f generally corresponds to the empirical risk of a loss function $\ell: \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_+$. More precisely, given n pairs of data points $\{(\boldsymbol{x}^{(i)}, y^{(i)}) \in \mathbb{R}^p \times \mathcal{Y};\ i = 1, \dots, n\}$, we have for linear models[4] $f(\boldsymbol{w}) := \frac{1}{n} \sum_{i=1}^n \ell(y^{(i)}, \boldsymbol{w}^\top \boldsymbol{x}^{(i)})$. Typical examples of differentiable loss functions are the square loss for least squares regression, i.e., $\ell(y, \hat{y}) = \frac{1}{2}(y - \hat{y})^2$ with y in $\mathbb{R}$, and the logistic loss $\ell(y, \hat{y}) = \log(1 + e^{-y\hat{y}})$ for logistic regression, with y in $\{-1, 1\}$. Clearly, several loss functions of interest are nondifferentiable, such as the hinge loss $\ell(y, \hat{y}) = (1 - y\hat{y})_+$ or the absolute deviation loss $\ell(y, \hat{y}) = |y - \hat{y}|$, for which most of the approaches we present in this monograph would not be applicable or require appropriate modifications. Given the tutorial character of this monograph, we restrict ourselves to smooth functions f, which we consider is a reasonably broad setting, and we refer the interested reader to appropriate references in Section 9. We refer the readers to [126] for a more complete description of loss functions.

Penalty or constraint? Given our convex data-fitting term $f(\boldsymbol{w})$, we consider in this monograph adding a convex penalty $\lambda\Omega(\boldsymbol{w})$. Within such a convex optimization framework, this is essentially equivalent to adding a constraint of the form $\Omega(\boldsymbol{w}) \leq \mu$. More precisely, under weak assumptions on f and Ω (on top of convexity), from Lagrange multiplier theory (see [20], Section 4.3) $\boldsymbol{w}$ is a solution of the constrained problem for a certain $\mu > 0$ if and only if it is a solution of the penalized problem for a certain $\lambda \geq 0$. Thus, the two regularization paths, i.e., the set of solutions when λ and μ vary, are equivalent. However, there is no direct mapping between corresponding values of λ and μ. Moreover, in a machine learning context, where the parameters λ and μ have to be selected, for example, through cross-validation, the penalized formulation tends to be empirically easier to tune, as the performance is

[4] In Section 1.5, we consider extensions to nonlinear predictors through multiple kernel learning.

usually quite robust to small changes in λ, while it is not robust to small changes in μ. Finally, we could also replace the penalization with a norm by a penalization with the squared norm. Indeed, following the same reasoning as for the nonsquared norm, a penalty of the form $\lambda\Omega(\boldsymbol{w})^2$ is "equivalent" to a constraint of the form $\Omega(\boldsymbol{w})^2 \leqslant \mu$, which itself is equivalent to $\Omega(\boldsymbol{w}) \leqslant \mu^{1/2}$, and thus to a penalty of the form $\lambda'\Omega(\boldsymbol{w})^2$, for $\lambda' \neq \lambda$. Thus, using a squared norm, as is often done in the context of multiple kernel learning (see Section 1.5), does not change the regularization properties of the formulation.

1.3 Sparsity-Inducing Norms

In this section, we present various norms as well as their main sparsity-inducing effects. These effects may be illustrated geometrically through the singularities of the corresponding unit balls (see Figure 1.4).

Sparsity through the ℓ_1-norm. When one knows *a priori* that the solutions $\boldsymbol{w}^\star$ of problem (1.1) should have a few nonzero coefficients, Ω is often chosen to be the ℓ_1-norm, i.e., $\Omega(\boldsymbol{w}) = \sum_{j=1}^{p} |\boldsymbol{w}_j|$. This leads for instance to the Lasso [133] or basis pursuit [37] with the square loss and to ℓ_1-regularized logistic regression (see, for instance, [75, 127]) with the logistic loss. Regularizing by the ℓ_1-norm is known to induce sparsity in the sense that, a number of coefficients of $\boldsymbol{w}^\star$, depending on the strength of the regularization, will be *exactly* equal to zero.

ℓ_1/ℓ_q-norms. In some situations, the coefficients of $\boldsymbol{w}^\star$ are naturally partitioned in subsets, or *groups*, of variables. This is typically the case, when working with ordinal variables.[5] It is then natural to select or remove *simultaneously* all the variables forming a group. A regularization norm exploiting explicitly this group structure, or ℓ_1-*group norm*, can be shown to improve the prediction performance and/or interpretability of the learned models [61, 83, 106, 116, 141, 156]. The

[5] Ordinal variables are integer-valued variables encoding levels of a certain feature, such as levels of severity of a certain symptom in a biomedical application, where the values do not correspond to an intrinsic linear scale: in that case it is common to introduce a vector of binary variables, each encoding a specific level of the symptom, that encodes collectively this single feature.

arguably simplest group norm is the so-called-ℓ_1/ℓ_2 norm:

$$\Omega(\boldsymbol{w}) := \sum_{g \in \mathcal{G}} d_g \|\boldsymbol{w}_g\|_2, \tag{1.2}$$

where $\mathcal{G}$ is a partition of $\{1,\dots,p\}$, $(d_g)_{g\in\mathcal{G}}$ are some strictly positive weights, and $\boldsymbol{w}_g$ denotes the vector in $\mathbb{R}^{|g|}$ recording the coefficients of $\boldsymbol{w}$ indexed by g in $\mathcal{G}$. Without loss of generality we may assume all weights $(d_g)_{g\in\mathcal{G}}$ to be equal to one (when $\mathcal{G}$ is a partition, we can rescale the components of $\boldsymbol{w}$ appropriately). As defined in Equation (1.2), Ω is known as a mixed ℓ_1/ℓ_2-norm. It behaves like an ℓ_1-norm on the vector $(\|\boldsymbol{w}_g\|_2)_{g\in\mathcal{G}}$ in $\mathbb{R}^{|\mathcal{G}|}$, and therefore, Ω induces group sparsity. In other words, each $\|\boldsymbol{w}_g\|_2$, and equivalently each $\boldsymbol{w}_g$, is encouraged to be set to zero. On the other hand, within the groups g in $\mathcal{G}$, the ℓ_2-norm does not promote sparsity. Combined with the square loss, it leads to the *group Lasso* formulation [141, 156]. Note that when $\mathcal{G}$ is the set of singletons, we retrieve the ℓ_1-norm. More general mixed ℓ_1/ℓ_q-norms for $q > 1$ are also used in the literature [157] (using $q = 1$ leads to a weighted ℓ_1-norm with no group-sparsity effects):

$$\Omega(\boldsymbol{w}) = \sum_{g\in\mathcal{G}} \|\boldsymbol{w}_g\|_q := \sum_{g\in\mathcal{G}} d_g \left\{ \sum_{j\in g} |\boldsymbol{w}_j|^q \right\}^{1/q}.$$

In practice though, the ℓ_1/ℓ_2- and ℓ_1/ℓ_∞-settings remain the most popular ones. Note that using ℓ_∞-norms may have the undesired effect to favor solutions $\boldsymbol{w}$ with many components of equal magnitude (due to the extra nondifferentiabilities away from zero). Grouped ℓ_1-norms are typically used when extra-knowledge is available regarding an appropriate partition, in particular in the presence of categorical variables with orthogonal encoding [116], for multi-task learning where joint variable selection is desired [106], and for multiple kernel learning (see Section 1.5).

Norms for overlapping groups: a direct formulation. In an attempt to better encode structural links between variables at play (e.g., spatial or hierarchical links related to the physics of the problem at hand), recent research has explored the setting where $\mathcal{G}$ in Equation (1.2) can contain groups of variables that *overlap* [9, 64, 66,

73, 121, 157]. In this case, if the groups span the entire set of variables, Ω is still a norm, and it yields sparsity in the form of specific patterns of variables. More precisely, the solutions $\boldsymbol{w}^\star$ of problem (1.1) can be shown to have a set of zero coefficients, or simply *zero pattern*, that corresponds to a union of some groups g in $\mathcal{G}$ [66]. This property makes it possible to control the sparsity patterns of $\boldsymbol{w}^\star$ by appropriately defining the groups in $\mathcal{G}$. Note that here the weights d_g should not be taken equal to one (see, e.g., [66] for more details). This form of *structured sparsity* has notably proven to be useful in various contexts, which we now illustrate through concrete examples:

— **One-dimensional sequence:** Given p variables organized in a sequence, if we want to select only contiguous nonzero patterns, we represent in Figure 1.1 the set of groups $\mathcal{G}$ to consider. In this case, we have $|\mathcal{G}| = O(p)$. Imposing the contiguity of the nonzero patterns is for instance relevant in the context of time series, or for the diagnosis of tumors, based on the profiles of arrayCGH [112]. Indeed, because of the specific spatial organization of bacterial artificial chromosomes along the genome, the set of discriminative features is expected to have specific contiguous patterns.

— **Two-dimensional grid:** In the same way, assume now that the p variables are organized on a two-dimensional grid. If we want the possible nonzero patterns $\mathcal{P}$ to be the set of all rectangles on this grid, the appropriate groups $\mathcal{G}$ to consider can be shown (see [66]) to be those represented in Figure 1.2. In this setting, we have $|\mathcal{G}| = O(\sqrt{p})$.

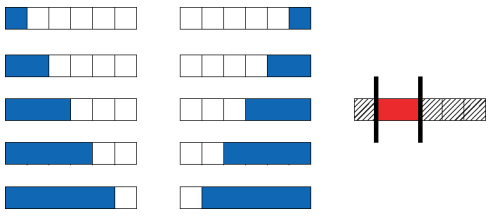


Fig. 1.1. (Left) The set of blue groups to penalize in order to select contiguous patterns in a sequence. (Right) In red, an example of such a nonzero pattern with its corresponding zero pattern (hatched area).

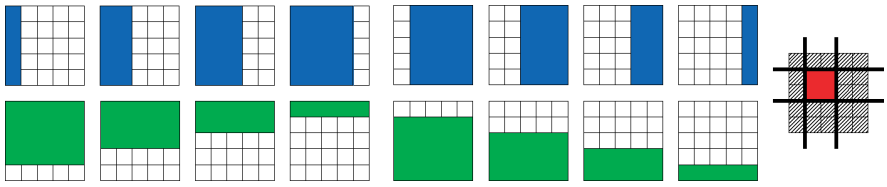


Fig. 1.2. Vertical and horizontal groups: (Left) the set of blue and green groups to penalize in order to select rectangles. (Right) In red, an example of nonzero pattern recovered in this setting, with its corresponding zero pattern (hatched area).

Sparsity-inducing regularizations built upon such group structures have resulted in good performances for background subtraction [62, 86, 88], topographic dictionary learning [72, 88], wavelet-based denoising [111], and for face recognition with corruption by occlusions [70].

— **Hierarchical structure:** A third interesting example assumes that the variables have a hierarchical structure. Specifically, we consider that the p variables correspond to the nodes of a tree $\mathcal{T}$ (or a forest of trees). Moreover, we assume that we want to select the variables according to a certain order: a feature can be selected only if all its ancestors in $\mathcal{T}$ are already selected. This hierarchical rule can be shown to lead to the family of groups displayed on Figure 1.3.
This resulting penalty was first used in [157]; since then, this group structure has led to numerous applications, for instance, wavelet-based denoising [15, 62, 69, 157], hierarchical dictionary learning for both topic modeling and image restoration [68, 69], log-linear models for the selection of potential orders of interaction in a probabilistic graphical model [121], bioinformatics, to exploit the tree structure of gene networks for multi-task regression [73], and multi-scale mining of fMRI data for the prediction of some cognitive task [67]. More recently, this hierarchical penalty was proved to be efficient for template selection in natural language processing [92].

— **Extensions:** The possible choices for the sets of groups $\mathcal{G}$ are not limited to the aforementioned examples. More

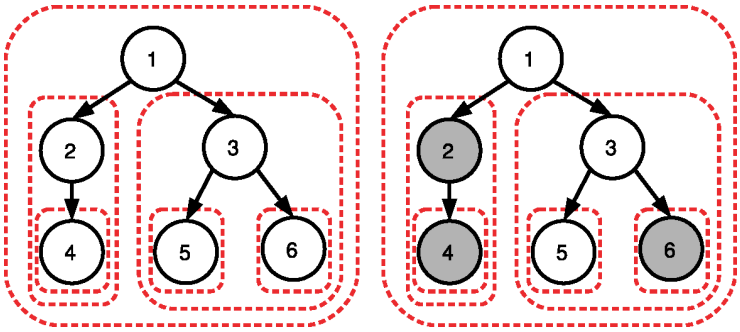


Fig. 1.3. Left: example of a tree-structured set of groups $\mathcal{G}$ (dashed contours in red), corresponding to a tree $\mathcal{T}$ with $p = 6$ nodes represented by black circles. Right: example of a sparsity pattern induced by the tree-structured norm corresponding to $\mathcal{G}$; the groups $\{2,4\},\{4\}$ and $\{6\}$ are set to zero, so that the corresponding nodes (in gray) that form subtrees of $\mathcal{T}$ are removed. The remaining nonzero variables $\{1,3,5\}$ form a rooted and connected subtree of $\mathcal{T}$. This sparsity pattern obeys the following equivalent rules: (i) if a node is selected, the same goes for all its ancestors; (ii) if a node is not selected, then its descendant are not selected.

complicated topologies can be considered, for instance, three-dimensional spaces discretized in cubes or spherical volumes discretized in slices; for instance, see [143] for an application to neuroimaging that pursues this idea. Moreover, directed acyclic graphs that extends the trees presented in Figure 1.3 have notably proven to be useful in the context of hierarchical variable selection [9, 121, 157],

Norms for overlapping groups: a latent variable formulation. The family of norms defined in Equation (1.2) is adapted to *intersection-closed* sets of nonzero patterns. However, some applications exhibit structures that can be more naturally modelled by *union-closed* families of supports. This idea was developed in [64, 105] where, given a set of groups $\mathcal{G}$, the following *latent group Lasso* norm was proposed:

$$\Omega_{\text{union}}(\boldsymbol{w}) := \min_{\boldsymbol{v}\in\mathbb{R}^{p\times|\mathcal{G}|}} \sum_{g\in\mathcal{G}} d_g \|\boldsymbol{v}^g\|_q, \quad \text{s.t.} \begin{cases} \sum_{g\in\mathcal{G}} \boldsymbol{v}^g = \boldsymbol{w}, \\ \forall g \in \mathcal{G},\ \boldsymbol{v}^g_j = 0 \quad \text{if } j \notin g. \end{cases}$$

The idea is to introduce latent parameter vectors $\boldsymbol{v}^g$ constrained each to be supported on the corresponding group g, which should explain $\boldsymbol{w}$

linearly and which are themselves regularized by a usual ℓ_1/ℓ_q-norm. Ω_{union} reduces to the usual ℓ_1/ℓ_q norm when groups are disjoint and provides therefore a different generalization of the latter to the case of overlapping groups than the norm considered in the previous paragraphs. In fact, it is easy to see that solving Equation (1.1) with the norm Ω_{union} is equivalent to solving

$$\min_{(\boldsymbol{v}^g \in \mathbb{R}^{|g|})_{g \in \mathcal{G}}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y^{(i)}, \sum_{g \in \mathcal{G}} \boldsymbol{v}_g^{g\top} \boldsymbol{x}_g^{(i)}\right) + \lambda \sum_{g \in \mathcal{G}} d_g \|\boldsymbol{v}^g\|_q \tag{1.3}$$

and setting $\boldsymbol{w} = \sum_{g \in \mathcal{G}} \boldsymbol{v}^g$. This last equation shows that using the norm Ω_{union} can be interpreted as implicitly duplicating the variables belonging to several groups and regularizing with a weighted ℓ_1/ℓ_q norm for disjoint groups in the expanded space. It should be noted that a careful choice of the weights is much more important in the situation of overlapping groups than in the case of disjoint groups, as it influences possible sparsity patterns [105].

This latent variable formulation pushes some of the vectors $\boldsymbol{v}^g$ to zero while keeping others with no zero components, hence leading to a vector $\boldsymbol{w}$ with a support which is in general the union of the selected groups. Interestingly, it can be seen as a convex relaxation of a non-convex penalty encouraging similar sparsity patterns which was introduced by [62]. Moreover, this norm can also be interpreted as a particular case of the family of *atomic norms*, which were recently introduced by [35].

Graph Lasso. One type of a priori knowledge commonly encountered takes the form of graph defined on the set of input variables, which is such that connected variables are more likely to be simultaneously relevant or irrelevant; this type of prior is common in genomics where regulation, co-expression or interaction networks between genes (or their expression level) used as predictors are often available. To favor the selection of neighbors of a selected variable, it is possible to consider the edges of the graph as groups in the previous formulation (see [64, 111]).

Patterns consisting of a small number of intervals. A quite similar situation occurs, when one knows a priori—typically for variables forming sequences (times series, strings, polymers)—that the support should consist of a small number of connected subsequences. In that case,

one can consider the sets of variables forming connected subsequences (or connected subsequences of length at most k) as the overlapping groups.

Multiple kernel learning. For most of the sparsity-inducing terms described in this monograph, we may replace real variables and their absolute values by pre-defined groups of variables with their Euclidean norms (we have already seen such examples with ℓ_1/ℓ_2-norms), or more generally, by members of reproducing kernel Hilbert spaces. As shown in Section 1.5, most of the tools that we present in this monograph are applicable to this case as well, through appropriate modifications and borrowing of tools from kernel methods. These tools have applications in particular in multiple kernel learning. Note that this extension requires tools from convex analysis presented in Section 1.4.

Trace norm. In learning problems on matrices, such as matrix completion, the rank plays a similar role to the cardinality of the support for vectors. Indeed, the rank of a matrix $\mathbf{M}$ may be seen as the number of non-zero singular values of $\mathbf{M}$. The rank of $\mathbf{M}$ however is not a continuous function of $\mathbf{M}$, and, following the convex relaxation of the ℓ_0-pseudo-norm into the ℓ_1-norm, we may relax the rank of $\mathbf{M}$ into the sum of its singular values, which happens to be a norm, and is often referred to as the trace norm or nuclear norm of $\mathbf{M}$, and which we denote by $\|\mathbf{M}\|_*$. As shown in this monograph, many of the tools designed for the ℓ_1-norm may be extended to the trace norm. Using the trace norm as a convex surrogate for rank has many applications in control theory [48], matrix completion [1, 130], multi-task learning [109], or multi-label classification [4], where low-rank priors are adapted.

Sparsity-inducing properties: A geometrical intuition. Although we consider in Equation (1.1) a regularized formulation, as already described in Section 1.2, we could equivalently focus on a *constrained* problem, that is,

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) \quad \text{such that} \quad \Omega(\boldsymbol{w}) \leq \mu, \tag{1.4}$$

for some $\mu \in \mathbb{R}_+$. The set of solutions of Equation (1.4) parameterized by μ is the same as that of Equation (1.1), as described by some value

of λ_μ depending on μ (e.g., see Section 3.2 in [20]). At optimality, the gradient of f evaluated at any solution $\hat{\boldsymbol{w}}$ of (1.4) is known to belong to the normal cone of $\mathcal{B} = \{\boldsymbol{w} \in \mathbb{R}^p;\ \Omega(\boldsymbol{w}) \leq \mu\}$ at $\hat{\boldsymbol{w}}$ [20]. In other words, for sufficiently small values of μ, i.e., so that the constraint is active, the level set of f for the value $f(\hat{\boldsymbol{w}})$ is tangent to $\mathcal{B}$.

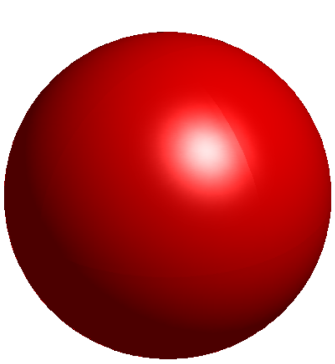
As a consequence, the geometry of the ball $\mathcal{B}$ is directly related to the properties of the solutions $\hat{\boldsymbol{w}}$. If Ω is taken to be the ℓ_2-norm, then the resulting ball $\mathcal{B}$ is the standard, isotropic, "round" ball that does not favor any specific direction of the space. On the other hand, when Ω is the ℓ_1-norm, $\mathcal{B}$ corresponds to a diamond-shaped pattern in two dimensions, and to a pyramid in three dimensions. In particular, $\mathcal{B}$ is anisotropic and exhibits some singular points due to the extra non-smoothness of Ω. Moreover, these singular points are located along the axis of $\mathbb{R}^p$, so that if the level set of f happens to be tangent at one of those points, sparse solutions are obtained. We display in Figure 1.4 the balls $\mathcal{B}$ for the ℓ_1-, ℓ_2-norms, and two different grouped ℓ_1/ℓ_2-norms.

Extensions. The design of sparsity-inducing norms is an active field of research and similar tools to the ones we present here can be derived for other norms. As shown in Section 3, computing the proximal operator readily leads to efficient algorithms, and for the extensions we present below, these operators can be efficiently computed.

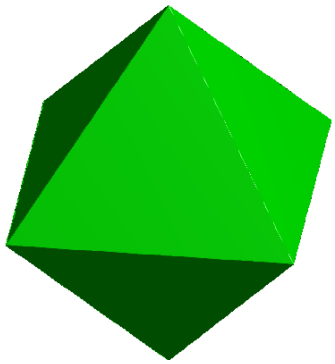
In order to impose prior knowledge on the support of predictor, the norms based on overlapping ℓ_1/ℓ_∞-norms can be shown to be convex relaxations of submodular functions of the support, and further ties can be made between convex optimization and combinatorial optimization (see [10] for more details). Moreover, similar developments may be carried through for norms which try to enforce that the predictors have many equal components and that the resulting clusters have specific shapes, e.g., contiguous in a pre-defined order, see some examples in Section 3, and, e.g., [11, 33, 86, 134, 144] and references therein.

1.4 Optimization Tools

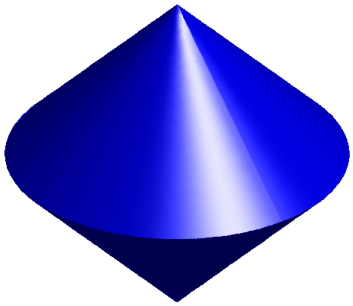
The tools used in this monograph are relatively basic and should be accessible to a broad audience. Most of them can be found in



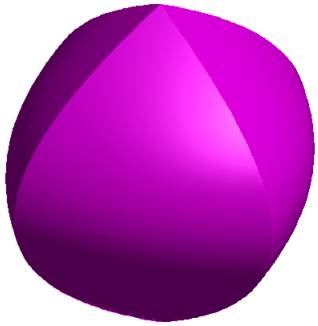
(a) ℓ_2-norm ball.



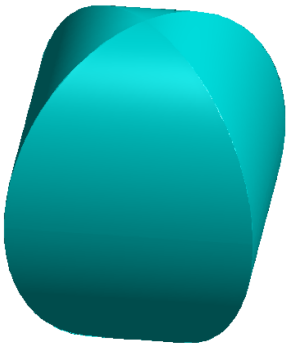
(b) ℓ_1-norm ball.



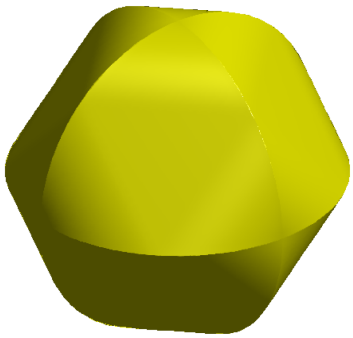
(c) ℓ_1/ℓ_2-norm ball: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}_{\{1,2\}}\|_2 + |\boldsymbol{w}_3|$.



(d) ℓ_1/ℓ_2-norm ball: $\Omega(\boldsymbol{w}) = \|\boldsymbol{w}\|_2 + |\boldsymbol{w}_1| + |\boldsymbol{w}_2|$.



(e) Ω_{union} ball for $\mathcal{G} = \big\{\{1,3\},\{2,3\}\big\}$.



(f) Ω_{union} ball for $\mathcal{G} = \big\{\{1,3\},\{2,3\},\{1,2\}\big\}$.

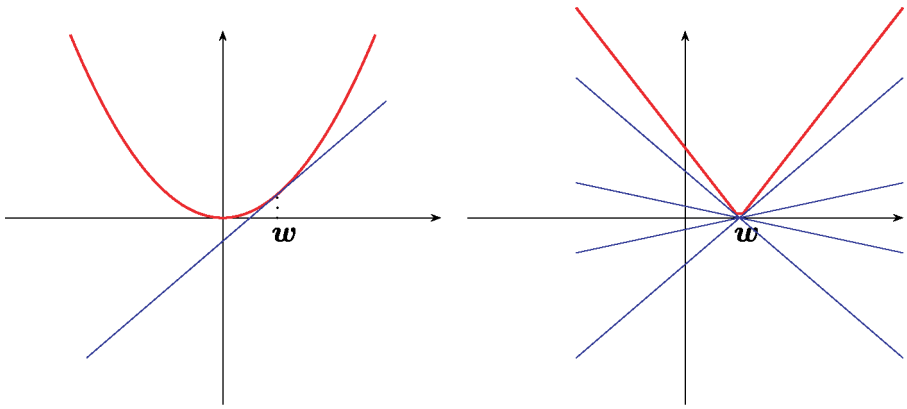
Fig. 1.4. Comparison between different balls of sparsity-inducing norms in three dimensions. The singular points appearing on these balls describe the sparsity-inducing behavior of the underlying norms Ω.

classical books on convex optimization [18, 20, 25, 104], but for self-containedness, we present here a few of them related to nonsmooth unconstrained optimization. In particular, these tools allow the derivation of rigorous approximate optimality conditions based on duality gaps (instead of relying on weak stopping criteria based on small changes or low-norm gradients).

Subgradients. Given a convex function $g: \mathbb{R}^p \to \mathbb{R}$ and a vector $\boldsymbol{w}$ in $\mathbb{R}^p$, let us define the *subdifferential* of g at $\boldsymbol{w}$ as

$$\partial g(\boldsymbol{w}) := \{\boldsymbol{z} \in \mathbb{R}^p \mid g(\boldsymbol{w}) + \boldsymbol{z}^\top(\boldsymbol{w}' - \boldsymbol{w}) \leq g(\boldsymbol{w}') \quad \text{for all vectors } \boldsymbol{w}' \in \mathbb{R}^p\}.$$

The elements of $\partial g(\boldsymbol{w})$ are called the *subgradients* of g at $\boldsymbol{w}$. Note that all convex functions defined on $\mathbb{R}^p$ have non-empty subdifferentials at every point. This definition admits a clear geometric interpretation: any subgradient $\boldsymbol{z}$ in $\partial g(\boldsymbol{w})$ defines an affine function $\boldsymbol{w}' \mapsto g(\boldsymbol{w}) + \boldsymbol{z}^\top(\boldsymbol{w}' - \boldsymbol{w})$ which is tangent to the graph of the function g (because of the convexity of g, it is a lower-bounding tangent). Moreover, there is a bijection (one-to-one correspondence) between such "tangent affine functions" and the subgradients, as illustrated in Figure 1.5.



(a) Smooth case. (b) Non-smooth case.

Fig. 1.5. Red curves represent the graph of a smooth (left) and a nonsmooth (right) function f. Blue affine functions represent subgradients of the function f at a point $\boldsymbol{w}$.

Subdifferentials are useful for studying nonsmooth optimization problems because of the following proposition (whose proof is straightforward from the definition):

Proposition 1.1 (Subgradients at Optimality).
For any convex function $g: \mathbb{R}^p \to \mathbb{R}$, a point $\boldsymbol{w}$ in $\mathbb{R}^p$ is a global minimum of g if and only if the condition $0 \in \partial g(\boldsymbol{w})$ holds.

Note that the concept of subdifferential is mainly useful for nonsmooth functions. If g is differentiable at $\boldsymbol{w}$, the set $\partial g(\boldsymbol{w})$ is indeed the singleton $\{\nabla g(\boldsymbol{w})\}$, where $\nabla g(\boldsymbol{w})$ is the gradient of g at $\boldsymbol{w}$, and the condition $0 \in \partial g(\boldsymbol{w})$ reduces to the classical first-order optimality condition $\nabla g(\boldsymbol{w}) = 0$. As a simple example, let us consider the following optimization problem

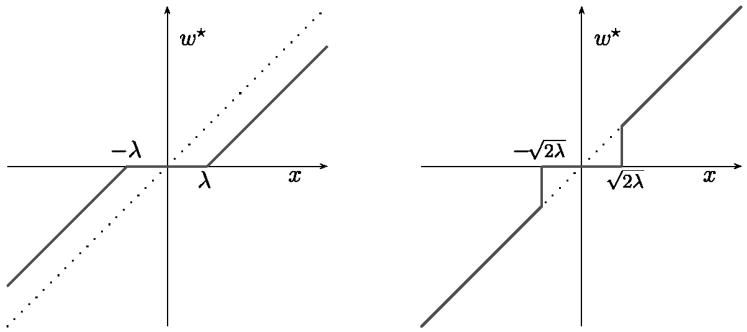
$$\min_{w \in \mathbb{R}} \frac{1}{2}(x - w)^2 + \lambda |w|.$$

Applying the previous proposition and noting that the subdifferential $\partial |\cdot|$ is $\{+1\}$ for $w > 0$, $\{-1\}$ for $w < 0$ and $[-1, 1]$ for $w = 0$, it is easy to show that the unique solution admits a closed form called the *soft-thresholding* operator, following a terminology introduced in [42]; it can be written

$$w^\star = \begin{cases} 0, & \text{if } |x| \leq \lambda \\ (1 - \frac{\lambda}{|x|})x, & \text{otherwise,} \end{cases} \tag{1.5}$$

or equivalently $w^\star = \text{sign}(x)(|x| - \lambda)_+$, where $\text{sign}(x)$ is equal to 1 if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. This operator is a core component of many optimization techniques for sparse estimation, as we shall see later. Its counterpart for nonconvex optimization problems is the hard-thresholding operator. Both of them are presented in Figure 1.6. Note that similar developments could be carried through using directional derivatives instead of subgradients (see, e.g., [20]).

Dual norm and optimality conditions. The next concept we introduce is the dual norm, which is important to study sparsity-inducing regularizations [9, 66, 99]. It notably arises in the analysis of estimation bounds [99], and in the design of working-set strategies



(a) soft-thresholding operator, $w^\star = \text{sign}(x)(|x| - \lambda)_+$, $\min_w \frac{1}{2}(x-w)^2 + \lambda|w|$.

(b) hard-thresholding operator $w^\star = \mathbf{1}_{|x| \geq \sqrt{2\lambda}} x$ $\min_w \frac{1}{2}(x-w)^2 + \lambda \mathbf{1}_{|w|>0}$.

Fig. 1.6. Soft- and hard-thresholding operators.

as will be shown in Section 6.1. The dual norm Ω^* of the norm Ω is defined for any vector $\boldsymbol{z}$ in $\mathbb{R}^p$ by

$$\Omega^*(\boldsymbol{z}) := \max_{\boldsymbol{w} \in \mathbb{R}^p} \boldsymbol{z}^\top \boldsymbol{w} \text{ such that } \Omega(\boldsymbol{w}) \leq 1. \tag{1.6}$$

Moreover, the dual norm of Ω^* is Ω itself, and as a consequence, the formula above holds also if the roles of Ω and Ω^* are exchanged. It is easy to show that in the case of an ℓ_q-norm, $q \in [1; +\infty]$, the dual norm is the $\ell_{q'}$-norm, with q' in $[1; +\infty]$ such that $\frac{1}{q} + \frac{1}{q'} = 1$. In particular, the ℓ_1- and ℓ_∞-norms are dual to each other, and the ℓ_2-norm is self-dual (dual to itself).

The dual norm plays a direct role in computing optimality conditions of sparse regularized problems. By applying Proposition 1.1 to Equation (1.1), we obtain the following proposition:

Proposition 1.2 (Optimality conditions for Equation (1.1)). Let us consider problem (1.1) where Ω is a norm on $\mathbb{R}^p$. A vector $\boldsymbol{w}$ in $\mathbb{R}^p$ is optimal if and only if $-\frac{1}{\lambda}\nabla f(\boldsymbol{w}) \in \partial\Omega(\boldsymbol{w})$ with

$$\partial\Omega(\boldsymbol{w}) = \begin{cases} \{\boldsymbol{z} \in \mathbb{R}^p;\ \Omega^*(\boldsymbol{z}) \leq 1\}, & \text{if } \boldsymbol{w} = 0, \\ \{\boldsymbol{z} \in \mathbb{R}^p;\ \Omega^*(\boldsymbol{z}) = 1 \text{ and } \boldsymbol{z}^\top \boldsymbol{w} = \Omega(\boldsymbol{w})\}, & \text{otherwise.} \end{cases} \tag{1.7}$$

Computing the subdifferential of a norm is a classical course exercise [20] and its proof will be presented in the next section, in Remark 1.1. As a consequence, the vector $\mathbf{0}$ is solution if and only if $\Omega^*\big(\nabla f(\mathbf{0})\big) \leq \lambda$. Note that this shows that for all λ larger than $\Omega^*\big(\nabla f(\mathbf{0})\big)$, $\boldsymbol{w} = \mathbf{0}$ is a solution of the regularized optimization problem (hence this value is the start of the non-trivial regularization path).

These general optimality conditions can be specialized to the Lasso problem [133], also known as basis pursuit [37]:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2 + \lambda\|\boldsymbol{w}\|_1, \tag{1.8}$$

where $\boldsymbol{y}$ is in $\mathbb{R}^n$, and $\boldsymbol{X}$ is a design matrix in $\mathbb{R}^{n\times p}$. With Equation (1.7) in hand, we can now derive necessary and sufficient optimality conditions:

Proposition 1.3 (Optimality conditions for the Lasso).
A vector $\boldsymbol{w}$ is a solution of the Lasso problem (1.8) if and only if

$$\forall j = 1,\dots,p, \quad \begin{cases} |\boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})| \leq n\lambda, & \text{if } \boldsymbol{w}_j = 0 \\ \boldsymbol{X}_j^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = n\lambda\operatorname{sign}(\boldsymbol{w}_j), & \text{if } \boldsymbol{w}_j \neq 0, \end{cases} \tag{1.9}$$

where $\boldsymbol{X}_j$ denotes the jth column of $\boldsymbol{X}$, and $\boldsymbol{w}_j$ the jth entry of $\boldsymbol{w}$.

Proof. We apply Proposition 1.2. The condition $-\frac{1}{\lambda}\nabla f(\boldsymbol{w}) \in \partial\|\boldsymbol{w}\|_1$ can be rewritten: $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) \in n\lambda\partial\|\boldsymbol{w}\|_1$, which is equivalent to: (i) if $\boldsymbol{w} = 0$, $\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_\infty \leq n\lambda$ (using the fact that the ℓ_∞-norm is dual to the ℓ_1-norm); (ii) if $\boldsymbol{w} \neq 0$, $\|\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w})\|_\infty = n\lambda$ and $\boldsymbol{w}^\top\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}) = n\lambda\|\boldsymbol{w}\|_1$. It is then easy to check that these conditions are equivalent to Equation (1.9). □

As we will see in Section 6.2, it is possible to derive from these conditions interesting properties of the Lasso, as well as efficient algorithms for solving it. We have presented a useful duality tool for norms. More generally, there exists a related concept for convex functions, which we now introduce.

1.4.1 Fenchel Conjugate and Duality Gaps

Let us denote by f^* the Fenchel conjugate of f [115], defined by

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

Fenchel conjugates are particularly useful to derive dual problems and duality gaps.[6] Under mild conditions, the conjugate of the conjugate of a convex function is itself, leading to the following representation of f as a maximum of affine functions:

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

In the context of this tutorial, it is notably useful to specify the expression of the conjugate of a norm. Perhaps surprisingly and misleadingly, the conjugate of a norm is not equal to its dual norm, but corresponds instead to the indicator function of the unit ball of its dual norm. More formally, let us introduce the indicator function ι_{Ω^*} such that $\iota_{\Omega^*}(\boldsymbol{z})$ is equal to 0 if $\Omega^*(\boldsymbol{z}) \leq 1$ and $+\infty$ otherwise. Then, we have the following well-known results, which appears in several text books (e.g., see Example 3.26 in [25]):

Proposition 1.4 (Fenchel conjugate of a norm). Let Ω be a norm on $\mathbb{R}^p$. The following equality holds for any $\boldsymbol{z} \in \mathbb{R}^p$

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

Proof. On the one hand, assume that the dual norm of $\boldsymbol{z}$ is greater than 1, that is, $\Omega^*(\boldsymbol{z}) > 1$. According to the definition of the dual norm (see Equation (1.6)), and since the supremum is taken over the compact set $\{\boldsymbol{w} \in \mathbb{R}^p;\ \Omega(\boldsymbol{w}) \leq 1\}$, there exists a vector $\boldsymbol{w}$ in this ball such that $\Omega^*(\boldsymbol{z}) = \boldsymbol{z}^\top \boldsymbol{w} > 1$. For any scalar $t \geq 0$, consider $\boldsymbol{v} = t\boldsymbol{w}$ and notice that

$$\boldsymbol{z}^\top \boldsymbol{v} - \Omega(\boldsymbol{v}) = t[\boldsymbol{z}^\top \boldsymbol{w} - \Omega(\boldsymbol{w})] \geq t,$$

[6] For many of our norms, *conic* duality tools would suffice (see, e.g., [25]).

which shows that when $\Omega^*(\boldsymbol{z}) > 1$, the Fenchel conjugate is unbounded. Now, assume that $\Omega^*(\boldsymbol{z}) \leq 1$. By applying the generalized Cauchy–Schwarz's inequality, we obtain for any $\boldsymbol{w}$

$$\boldsymbol{z}^\top \boldsymbol{w} - \Omega(\boldsymbol{w}) \leq \Omega^*(\boldsymbol{z})\Omega(\boldsymbol{w}) - \Omega(\boldsymbol{w}) \leq 0.$$

Equality holds for $\boldsymbol{w} = \boldsymbol{0}$, and the conclusion follows. □

An important and useful duality result is the so-called Fenchel–Young inequality (see [20]), which we will shortly illustrate geometrically:

Proposition 1.5 (Fenchel–Young inequality). Let $\boldsymbol{w}$ be a vector in $\mathbb{R}^p$, f be a function on $\mathbb{R}^p$, and $\boldsymbol{z}$ be a vector in the domain of f^* (which we assume non-empty). We have then the following inequality

$$f(\boldsymbol{w}) + f^*(\boldsymbol{z}) \geq \boldsymbol{w}^\top \boldsymbol{z},$$

with equality if and only if $\boldsymbol{z}$ is in $\partial f(\boldsymbol{w})$.

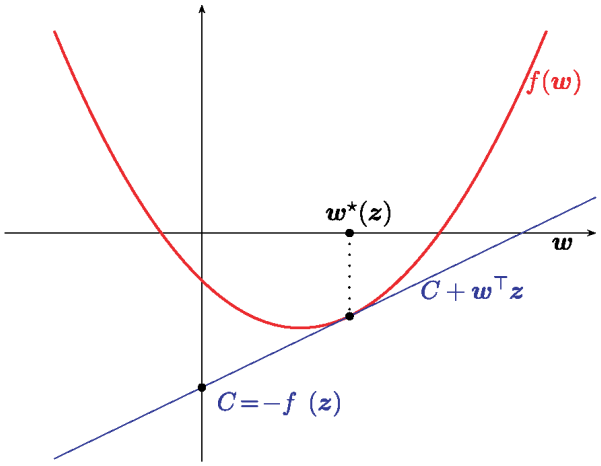
We can now illustrate geometrically the duality principle between a function and its Fenchel conjugate in Figure 1.7.

Remark 1.1. With Proposition 1.4 in place, we can formally (and easily) prove the relationship in Equation (1.7) that make explicit the subdifferential of a norm. Based on Proposition 1.4, we indeed know that the conjugate of Ω is ι_{Ω^*}. Applying the Fenchel–Young inequality (Proposition 1.5), we have

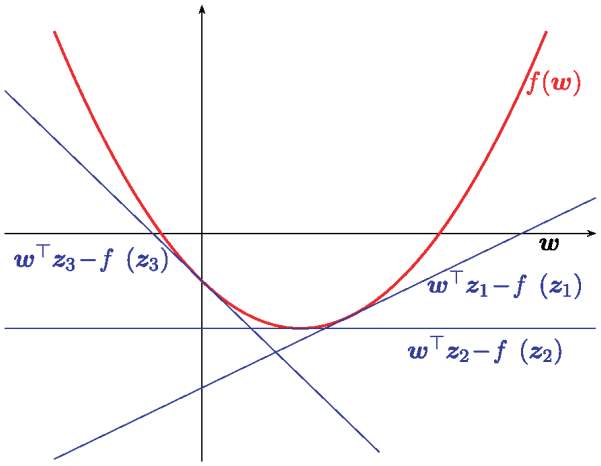
$$\boldsymbol{z} \in \partial\Omega(\boldsymbol{w}) \Leftrightarrow \left[\boldsymbol{z}^\top \boldsymbol{w} = \Omega(\boldsymbol{w}) + \iota_{\Omega^*}(\boldsymbol{z})\right],$$

which leads to the desired conclusion.

For many objective functions, the Fenchel conjugate admits closed forms, and can therefore be computed efficiently [20]. Then, it is



(a) Fenchel conjugate, tangent hyperplanes and subgradients.



(b) The graph of f is the envelope of the tangent hyperplanes $\mathcal{P}(z)$.

Fig. 1.7. For all z in $\mathbb{R}^p$, we denote by $\mathcal{P}(z)$ the hyperplane with normal z and tangent to the graph of the convex function f. (a) For any contact point between the graph of f and an hyperplane $\mathcal{P}(z)$, we have that $f(w) + f^*(z) = w^\top z$ and z is in $\partial f(w)$ (the Fenchel–Young inequality is an equality). (b) The graph of f is the convex envelope of the collection of hyperplanes $(\mathcal{P}(z))_{z \in \mathbb{R}^p}$.

possible to derive a duality gap for problem (1.1) from standard Fenchel duality arguments (see [20]), as shown in the following proposition:

Proposition 1.6 (Duality for Problem (1.1)). If f^* and Ω^* are respectively, the Fenchel conjugate of a convex and differentiable function f and the dual norm of Ω, then we have

$$\max_{\boldsymbol{z}\in\mathbb{R}^p:\Omega^*(\boldsymbol{z})\leq\lambda} -f^*(\boldsymbol{z}) \leq \min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w}). \tag{1.10}$$

Moreover, equality holds as soon as the domain of f has non-empty interior.

Proof. This result is a specific instance of Theorem 3.3.5 in [20]. In particular, we use the fact that the conjugate of a norm Ω is the indicator function ι_{Ω^*} of the unit ball of the dual norm Ω^* (see Proposition 1.4). □

If $\boldsymbol{w}^\star$ is a solution of Equation (1.1), and $\boldsymbol{w}, \boldsymbol{z}$ in $\mathbb{R}^p$ are such that $\Omega^*(\boldsymbol{z}) \leq \lambda$, this proposition implies that we have

$$f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w}) \geq f(\boldsymbol{w}^\star) + \lambda\Omega(\boldsymbol{w}^\star) \geq -f^*(\boldsymbol{z}). \tag{1.11}$$

The difference between the left and right term of Equation (1.11) is called a duality gap. It represents the difference between the value of the primal objective function $f(\boldsymbol{w}) + \lambda\Omega(\boldsymbol{w})$ and a dual objective function $-f^*(\boldsymbol{z})$, where $\boldsymbol{z}$ is a dual variable. The proposition says that the duality gap for a pair of optima $\boldsymbol{w}^\star$ and $\boldsymbol{z}^\star$ of the primal and dual problem is equal to 0. When the optimal duality gap is zero one says that *strong duality* holds. In our situation, the duality gap for the pair of primal/dual problems in Equation (1.10), may be decomposed as the sum of two non-negative terms (as the consequence of Fenchel–Young inequality):

$$(f(\boldsymbol{w}) + f^*(\boldsymbol{z}) - \boldsymbol{w}^\top\boldsymbol{z}) + \lambda(\Omega(\boldsymbol{w}) + \boldsymbol{w}^\top(\boldsymbol{z}/\lambda) + \iota_{\Omega^*}(\boldsymbol{z}/\lambda)).$$

It is equal to zero if and only if the two terms are simultaneously equal to zero.

Duality gaps are important in convex optimization because they provide an upper bound on the difference between the current value of

an objective function and the optimal value, which makes it possible to set proper stopping criteria for iterative optimization algorithms. Given a current iterate $\boldsymbol{w}$, computing a duality gap requires choosing a "good" value for $\boldsymbol{z}$ (and in particular a feasible one). Given that at optimality, $\boldsymbol{z}(\boldsymbol{w}^\star) = \nabla f(\boldsymbol{w}^\star)$ is the unique solution to the dual problem, a natural choice of dual variable is $\boldsymbol{z} = \min\big(1, \frac{\lambda}{\Omega^*(\nabla f(\boldsymbol{w}))}\big)\nabla f(\boldsymbol{w})$, which reduces to $\boldsymbol{z}(\boldsymbol{w}^\star)$ at the optimum and therefore yields a zero duality gap at optimality.

Note that in most formulations that we will consider, the function f is of the form $f(\boldsymbol{w}) = \psi(\boldsymbol{X}\boldsymbol{w})$ with $\psi{:}\mathbb{R}^n \to \mathbb{R}$ and $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ a design matrix. Indeed, this corresponds to linear prediction on $\mathbb{R}^p$, given n observations $\boldsymbol{x}_i$, $i = 1,\dots,n$, and the predictions $\boldsymbol{X}\boldsymbol{w} = (\boldsymbol{w}^\top \boldsymbol{x}_i)_{i=1,\dots,n}$. Typically, the Fenchel conjugate of ψ is easy to compute[7] while the design matrix $\boldsymbol{X}$ makes it hard[8] to compute f^*. In that case, Equation (1.1) can be rewritten as

$$\min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{u}\in\mathbb{R}^n} \psi(\boldsymbol{u}) + \lambda\, \Omega(\boldsymbol{w}), \quad \text{s.t. } \boldsymbol{u} = \boldsymbol{X}\boldsymbol{w}, \tag{1.12}$$

and equivalently as the optimization of the Lagrangian

$$\begin{aligned} &\min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{u}\in\mathbb{R}^n} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \psi(\boldsymbol{u}) + \lambda\Omega(\boldsymbol{w}) + \lambda\boldsymbol{\alpha}^\top(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{u}), \\ &\min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{u}\in\mathbb{R}^n} \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} (\psi(\boldsymbol{u}) - \lambda\boldsymbol{\alpha}^\top\boldsymbol{u}) + \lambda(\Omega(\boldsymbol{w}) + \boldsymbol{\alpha}^\top\boldsymbol{X}\boldsymbol{w}), \end{aligned} \tag{1.13}$$

which is obtained by introducing the Lagrange multiplier $\boldsymbol{\alpha}$ for the constraint $\boldsymbol{u} = \boldsymbol{X}\boldsymbol{w}$. The corresponding Fenchel dual[9] is then

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\psi^*(\lambda\boldsymbol{\alpha}) \quad \text{such that } \Omega^*(\boldsymbol{X}^\top\boldsymbol{\alpha}) \le 1, \tag{1.14}$$

which does not require any inversion of $\boldsymbol{X}^\top\boldsymbol{X}$ (which would be required for computing the Fenchel conjugate of f). Thus, given a candidate $\boldsymbol{w}$, we consider $\boldsymbol{\alpha} = \min\big(1, \frac{\lambda}{\Omega^*(\boldsymbol{X}^\top\nabla\psi(\boldsymbol{X}\boldsymbol{w}))}\big)\nabla\psi(\boldsymbol{X}\boldsymbol{w})$, and can get

[7] For the least-squares loss with output vector $\boldsymbol{y} \in \mathbb{R}^n$, we have $\psi(\boldsymbol{u}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{u}\|_2^2$ and $\psi^*(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{\beta}\|_2^2 + \boldsymbol{\beta}^\top\boldsymbol{y}$. For the logistic loss, we have $\psi(\boldsymbol{u}) = \sum_{i=1}^n \log(1 + \exp(-\boldsymbol{y}_i\boldsymbol{u}_i))$ and $\psi^*(\boldsymbol{\beta}) = \sum_{i=1}^n (1 + \boldsymbol{\beta}_i\boldsymbol{y}_i)\log(1 + \boldsymbol{\beta}_i\boldsymbol{y}_i) - \boldsymbol{\beta}_i\boldsymbol{y}_i\log(-\boldsymbol{\beta}_i\boldsymbol{y}_i)$ if $\forall i,\ -\boldsymbol{\beta}_i\boldsymbol{y}_i \in [0,1]$ and $+\infty$ otherwise.

[8] It would require to compute the pseudo-inverse of $\boldsymbol{X}$.

[9] Fenchel conjugacy naturally extends to this case; see Theorem 3.3.5 in [20] for more details.

an upper bound on optimality using primal (1.12) and dual (1.14) problems. Concrete examples of such duality gaps for various sparse regularized problems are presented in Appendix D of [85], and are implemented in the open-source software SPAMS,[10] which we have used in the experimental section of this monograph.

1.4.2 Quadratic Variational Formulation of Norms

Several variational formulations are associated with norms, the most natural one being the one that results directly from (1.6) applied to the dual norm:

$$\Omega(\boldsymbol{w}) = \max_{\boldsymbol{z} \in \mathbb{R}^p} \boldsymbol{w}^\top \boldsymbol{z} \quad \text{s.t.} \quad \Omega^*(\boldsymbol{z}) \leq 1.$$

However, another type of variational form is quite useful, especially for sparsity-inducing norms; among other purposes, as it is obtained by a variational upper-bound (as opposed to a lower-bound in the equation above), it leads to a general algorithmic scheme for learning problems regularized with this norm, in which the difficulties associated with optimizing the loss and that of optimizing the norm are partially decoupled. We present it in Section 5. We introduce this variational form first for the ℓ_1- and ℓ_1/ℓ_2-norms and subsequently generalize it to norms that we call *subquadratic norms*.

The case of the ℓ_1- and ℓ_1/ℓ_2-norms. The two basic variational identities we use are, for $a, b > 0$,

$$2ab = \inf_{\eta \in \mathbb{R}_+^*} \eta^{-1} a^2 + \eta b^2, \tag{1.15}$$

where the infimum is attained at $\eta = a/b$, and, for $\boldsymbol{a} \in \mathbb{R}_+^p$,

$$\left(\sum_{i=1}^p \boldsymbol{a}_i \right)^2 = \inf_{\boldsymbol{\eta} \in (\mathbb{R}_+^*)^p} \sum_{i=1}^p \frac{\boldsymbol{a}_i^2}{\boldsymbol{\eta}_i} \text{ s.t. } \sum_{i=1}^p \boldsymbol{\eta}_i = 1. \tag{1.16}$$

The last identity is a direct consequence of the Cauchy–Schwarz inequality:

$$\sum_{i=1}^p \boldsymbol{a}_i = \sum_{i=1}^p \frac{\boldsymbol{a}_i}{\sqrt{\boldsymbol{\eta}_i}} \cdot \sqrt{\boldsymbol{\eta}_i} \leq \left(\sum_{i=1}^p \frac{\boldsymbol{a}_i^2}{\boldsymbol{\eta}_i} \right)^{1/2} \left(\sum_{i=1}^p \boldsymbol{\eta}_i \right)^{1/2}. \tag{1.17}$$

[10] http://www.di.ens.fr/willow/SPAMS/.

The infima in the previous expressions can be replaced by a minimization if the function $q\colon \mathbb{R} \times \mathbb{R}_+ \to \mathbb{R}_+$ with $q(x,y) = \frac{x^2}{y}$ is extended in $(0,0)$ using the convention "0/0=0", since the resulting function[11] is a proper closed convex function. We will use this convention implicitly from now on. The minimum is then attained when equality holds in the Cauchy–Schwarz inequality, that is for $\sqrt{\boldsymbol{\eta}_i} \propto \boldsymbol{a}_i/\sqrt{\boldsymbol{\eta}_i}$, which leads to $\boldsymbol{\eta}_i = \frac{\boldsymbol{a}_i}{\|\boldsymbol{a}\|_1}$ if $\boldsymbol{a} \neq 0$ and 0 else.

Introducing the simplex $\triangle_p = \{\boldsymbol{\eta} \in \mathbb{R}^p_+ \mid \sum_{i=1}^p \boldsymbol{\eta}_i = 1\}$, we apply these variational forms to the ℓ_1- and ℓ_1/ℓ_2-norms (with nonoverlapping groups) with $\|\boldsymbol{w}\|_{\ell_1/\ell_2} = \sum_{g\in\mathcal{G}} \|\boldsymbol{w}_g\|_2$ and $|\mathcal{G}| = m$, so that we obtain directly:

$$\|\boldsymbol{w}\|_1 = \min_{\boldsymbol{\eta}\in\mathbb{R}^p_+} \frac{1}{2}\sum_{i=1}^p \left[\frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} + \boldsymbol{\eta}_i\right], \qquad \|\boldsymbol{w}\|_1^2 = \min_{\boldsymbol{\eta}\in\triangle_p} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i},$$

$$\|\boldsymbol{w}\|_{\ell_1/\ell_2} = \min_{\boldsymbol{\eta}\in\mathbb{R}^m_+} \frac{1}{2}\sum_{g\in\mathcal{G}} \left[\frac{\|\boldsymbol{w}_g\|_2^2}{\boldsymbol{\eta}^g} + \boldsymbol{\eta}^g\right], \quad \|\boldsymbol{w}\|_{\ell_1/\ell_2}^2 = \min_{\boldsymbol{\eta}\in\triangle_m} \sum_{g\in\mathcal{G}} \frac{\|\boldsymbol{w}_g\|_2^2}{\boldsymbol{\eta}^g}.$$

Quadratic variational forms for subquadratic norms. The variational form of the ℓ_1-norm admits a natural generalization for certain norms that we call *subquadratic* norms. Before we introduce them, we review a few useful properties of norms. In this section, we will denote $|\boldsymbol{w}|$ the vector $(|\boldsymbol{w}_1|,\dots,|\boldsymbol{w}_p|)$.

Definition 1.1 (Absolute and monotonic norm). We say that:

- A norm Ω is **absolute** if for all $v \in \mathbb{R}^p$, $\Omega(\boldsymbol{v}) = \Omega(|\boldsymbol{v}|)$.
- A norm Ω is **monotonic** if for all $\boldsymbol{v},\boldsymbol{w} \in \mathbb{R}^p$ s.t. $|\boldsymbol{v}_i| \leq |\boldsymbol{w}_i|$, $i = 1,\dots,p$, it holds that $\Omega(\boldsymbol{v}) \leq \Omega(\boldsymbol{w})$.

These definitions are in fact equivalent (see, e.g., [16]):

Proposition 1.7. A norm is *monotonic* if and only if it is *absolute*.

---

[11] This extension is in fact the function $\tilde{q}\colon (x,y) \mapsto \min\left\{t \in \mathbb{R}_+ \mid \begin{bmatrix} t & x \\ x & y \end{bmatrix} \succeq 0\right\}$.

Proof. If Ω is monotonic, the fact that $\left|\boldsymbol{v}\right| = \left||\boldsymbol{v}|\right|$ implies $\Omega(\boldsymbol{v}) = \Omega(|\boldsymbol{v}|)$ so that Ω is absolute.

If Ω is absolute, we first show that Ω^* is absolute. Indeed,

$$\Omega^*(\boldsymbol{\kappa}) = \max_{\boldsymbol{w}\in\mathbb{R}^p,\, \Omega(|\boldsymbol{w}|)\leq 1} \boldsymbol{w}^\top\boldsymbol{\kappa} = \max_{\boldsymbol{w}\in\mathbb{R}^p,\, \Omega(|\boldsymbol{w}|)\leq 1} |\boldsymbol{w}|^\top|\boldsymbol{\kappa}| = \Omega^*(|\boldsymbol{\kappa}|).$$

Then if $|\boldsymbol{v}| \leq |\boldsymbol{w}|$, since $\Omega^*(\boldsymbol{\kappa}) = \Omega^*(|\boldsymbol{\kappa}|)$,

$$\Omega(\boldsymbol{v}) = \max_{\boldsymbol{\kappa}\in\mathbb{R}^p,\, \Omega^*(|\boldsymbol{\kappa}|)\leq 1} |\boldsymbol{v}|^\top|\boldsymbol{\kappa}| \leq \max_{\boldsymbol{\kappa}\in\mathbb{R}^p,\, \Omega^*(|\boldsymbol{\kappa}|)\leq 1} |\boldsymbol{w}|^\top|\boldsymbol{\kappa}| = \Omega(\boldsymbol{w}),$$

which shows that Ω is monotonic. □

We now introduce a family of norms, which have recently been studied in [93].

Definition 1.2 (H-norm). Let H be a compact convex subset of $\mathbb{R}^p_+$, such that $H \cap (\mathbb{R}^*_+)^p \neq \varnothing$, we say that Ω_H is an H-norm if $\Omega_H(\boldsymbol{w}) = \min_{\boldsymbol{\eta}\in H} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i}$.

The next proposition shows that Ω_H is indeed a norm and characterizes its dual norm.

Proposition 1.8. Ω_H is a norm and $\Omega_H^*(\boldsymbol{\kappa})^2 = \max_{\boldsymbol{\eta}\in H} \sum_{i=1}^p \boldsymbol{\eta}_i \boldsymbol{\kappa}_i^2$.

Proof. First, since H contains at least one element whose components are all strictly positive, Ω is finite on $\mathbb{R}^p$. Symmetry, nonnegativity and homogeneity of Ω_H are straightforward from the definitions. Definiteness results from the fact that H is bounded. Ω_H is convex, since it is obtained by minimization of $\boldsymbol{\eta}$ in a jointly convex formulation. Thus Ω_H is a norm. Finally,

$$\begin{aligned}\frac{1}{2}\Omega_H^*(\boldsymbol{\kappa})^2 &= \max_{\boldsymbol{w}\in\mathbb{R}^p} \boldsymbol{w}^\top\boldsymbol{\kappa} - \frac{1}{2}\Omega_H(\boldsymbol{w})^2 \\ &= \max_{\boldsymbol{w}\in\mathbb{R}^p} \max_{\boldsymbol{\eta}\in H} \boldsymbol{w}^\top\boldsymbol{\kappa} - \frac{1}{2}\boldsymbol{w}^\top \mathrm{Diag}(\boldsymbol{\eta})^{-1}\boldsymbol{w}.\end{aligned}$$

The form of the dual norm follows by maximizing w.r.t. $\boldsymbol{w}$. □

We finally introduce the family of norms that we call *subquadratic*.

Definition 1.3 (Subquadratic norm). Let Ω and Ω^* a pair of *absolute* dual norms. Let $\bar{\Omega}^*$ be the function defined as $\bar{\Omega}^*\colon \boldsymbol{\kappa} \mapsto [\Omega^*(|\boldsymbol{\kappa}|^{1/2})]^2$ where we use the notation $|\boldsymbol{\kappa}|^{1/2} = (|\boldsymbol{\kappa}_1|^{1/2}, \dots, |\boldsymbol{\kappa}_p|^{1/2})^\top$. We say that Ω is *subquadratic* if $\bar{\Omega}^*$ is convex.

With this definition, we have:

Lemma 1.9. If Ω is *subquadratic*, then $\bar{\Omega}^*$ is a norm, and denoting $\bar{\Omega}$ the dual norm of the latter, we have:

$$\Omega(\boldsymbol{w}) = \frac{1}{2} \min_{\boldsymbol{\eta} \in \mathbb{R}^p_+} \sum_i \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} + \bar{\Omega}(\boldsymbol{\eta})$$

$$\Omega(\boldsymbol{w})^2 = \min_{\boldsymbol{\eta} \in H} \sum_i \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} \quad \text{where } H = \{\boldsymbol{\eta} \in \mathbb{R}^p_+ \mid \bar{\Omega}(\boldsymbol{\eta}) \leq 1\}.$$

Proof. Note that by construction, $\bar{\Omega}^*$ is homogeneous, symmetric and definite ($\bar{\Omega}^*(\boldsymbol{\kappa}) = 0 \Rightarrow \boldsymbol{\kappa} = 0$). If $\bar{\Omega}^*$ is convex then $\bar{\Omega}^*(\frac{1}{2}(\boldsymbol{v} + \boldsymbol{u})) \leq \frac{1}{2}\big(\bar{\Omega}^*(\boldsymbol{v}) + \bar{\Omega}^*(\boldsymbol{u})\big)$, which by homogeneity shows that $\bar{\Omega}^*$ also satisfies the triangle inequality. Together, these properties show that $\bar{\Omega}^*$ is a norm. To prove the first identity we have, applying (1.15), and since Ω is absolute,

$$\begin{aligned}
\Omega(\boldsymbol{w}) &= \max_{\boldsymbol{\kappa} \in \mathbb{R}^p_+} \boldsymbol{\kappa}^\top |\boldsymbol{w}|, \quad \text{s.t.} \quad \Omega^*(\boldsymbol{\kappa}) \leq 1 \\
&= \max_{\boldsymbol{\kappa} \in \mathbb{R}^p_+} \sum_{i=1}^p \boldsymbol{\kappa}_i^{1/2} |\boldsymbol{w}_i|, \quad \text{s.t.} \quad \Omega^*(\boldsymbol{\kappa}^{1/2})^2 \leq 1 \\
&= \max_{\boldsymbol{\kappa} \in \mathbb{R}^p_+} \min_{\boldsymbol{\eta} \in \mathbb{R}^p_+} \frac{1}{2} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} + \boldsymbol{\kappa}^\top \boldsymbol{\eta}, \quad \text{s.t.} \quad \bar{\Omega}^*(\boldsymbol{\kappa}) \leq 1 \\
&= \min_{\boldsymbol{\eta} \in \mathbb{R}^p_+} \max_{\boldsymbol{\kappa} \in \mathbb{R}^p_+} \frac{1}{2} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} + \boldsymbol{\kappa}^\top \boldsymbol{\eta}, \quad \text{s.t.} \quad \bar{\Omega}^*(\boldsymbol{\kappa}) \leq 1,
\end{aligned}$$

which proves the first variational formulation (note that we can switch the order of the max and min operations because strong duality holds,

which is due to the non-emptiness of the unit ball of the dual norm). The second one follows similarly by applying (1.16) instead of (1.15).

$$\begin{aligned}\Omega(\boldsymbol{w})^2 &= \max_{\boldsymbol{\kappa}\in\mathbb{R}^p_+} \left(\sum_{i=1}^p \boldsymbol{\kappa}_i^{1/2}|\boldsymbol{w}_i|\right)^2, \quad \text{s.t.} \quad \Omega^*(\boldsymbol{\kappa}^{1/2})^2 \leq 1 \\ &= \max_{\boldsymbol{\kappa}\in\mathbb{R}^p_+} \min_{\tilde{\boldsymbol{\eta}}\in\mathbb{R}^p_+} \sum_{i=1}^p \frac{\boldsymbol{\kappa}_i \boldsymbol{w}_i^2}{\tilde{\boldsymbol{\eta}}_i}, \quad \text{s.t.} \quad \sum_{i=1}^p \tilde{\boldsymbol{\eta}}_i = 1,\ \bar{\Omega}^*(\boldsymbol{\kappa}) \leq 1 \\ &= \max_{\boldsymbol{\kappa}\in\mathbb{R}^p_+} \min_{\boldsymbol{\eta}\in\mathbb{R}^p_+} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i}, \quad \text{s.t.} \quad \boldsymbol{\eta}^\top \boldsymbol{\kappa} = 1,\ \bar{\Omega}^*(\boldsymbol{\kappa}) \leq 1. \qquad \square\end{aligned}$$

Thus, given a subquadratic norm, we may define a convex set H, namely the intersection of the unit ball of $\bar{\Omega}$ with the positive orthant $\mathbb{R}^p_+$, such that $\Omega(\boldsymbol{w})^2 = \min_{\boldsymbol{\eta}\in H} \sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i}$, i.e., a subquadratic norm is an H-norm. We now show that these two properties are in fact equivalent.

Proposition 1.10. Ω is *subquadratic* if and only if it is an H-norm.

Proof. The previous lemma shows that subquadratic norms are H-norms. Conversely, let Ω_H be an H-norm. By construction, Ω_H is absolute, and as a result of Proposition 1.8, $\bar{\Omega}_H^*(\boldsymbol{w}) = \left(\Omega_H^*(|\boldsymbol{w}|^{1/2})\right)^2 = \max_{\boldsymbol{\eta}\in H} \sum_i \boldsymbol{\eta}_i |\boldsymbol{w}_i|$, which shows that $\bar{\Omega}_H^*$ is a convex function, as a maximum of convex functions. $\square$

It should be noted that the set H leading to a given H-norm Ω_H is not unique; in particular H is not necessarily the intersection of the unit ball of a norm with the positive orthant. Indeed, for the ℓ_1-norm, we can take H to be the unit simplex.

Proposition 1.11. Given a convex compact set H, let Ω_H be the associated H-norm and $\bar{\Omega}_H$ as defined in Lemma 1.9. Define the mirror image of H as the set $\mathrm{Mirr}(H) = \{\boldsymbol{v} \in \mathbb{R}^p \mid |\boldsymbol{v}| \in H\}$ and denote the convex hull of a set S by $\mathrm{Conv}(S)$. Then the unit ball of $\bar{\Omega}_H$ is $\mathrm{Conv}(\mathrm{Mirr}(H))$.

Proof. By construction:

$$\bar{\Omega}_H^*(\boldsymbol{\kappa}) = \Omega_H^*(|\boldsymbol{\kappa}|^{1/2})^2 = \max_{\boldsymbol{\eta}\in H} \boldsymbol{\eta}^\top |\boldsymbol{\kappa}|$$
$$= \max_{|\boldsymbol{w}|\in H} \boldsymbol{w}^\top \boldsymbol{\kappa} = \max_{\boldsymbol{w}\in \mathrm{Conv}(\mathrm{Mirr}(H))} \boldsymbol{w}^\top \boldsymbol{\kappa},$$

since the maximum of a convex function over a convex set is attained at its extreme points. But $C = \mathrm{Conv}(\mathrm{Mirr}(H))$ is by construction a centrally symmetric convex set, which is bounded and closed like H, and whose interior contains 0 since H contains at least one point whose components are strictly positive. This implies by Theorem 15.2 in [115] that C is the unit ball of a norm (namely $\boldsymbol{x} \mapsto \inf\{\lambda \in \mathbb{R}_+ \mid \boldsymbol{x} \in \lambda C\}$), which by duality has to be the unit ball of $\bar{\Omega}_H$. □

This proposition combined with the result of Lemma 1.9 therefore shows that if $\mathrm{Conv}(\mathrm{Mirr}(H)) = \mathrm{Conv}(\mathrm{Mirr}(H'))$ then H and H' define the same norm.

Several instances of the general variational form we considered in this section have appeared in the literature [70, 109, 110]. For norms that are not subquadratic, it is often the case that their dual norm is itself subquadratic, in which case symmetric variational forms can be obtained [2]. Finally, we show in Section 5 that all norms admit a quadratic variational form provided the bilinear form considered is allowed to be non-diagonal.

1.5 Multiple Kernel Learning

A seemingly unrelated problem in machine learning, the problem of *multiple kernel learning* is in fact intimately connected with sparsity-inducing norms by duality. It actually corresponds to the most natural extension of sparsity to reproducing kernel Hilbert spaces. We will show that for a large class of norms and, among them, many sparsity-inducing norms, there exists for each of them a corresponding multiple kernel learning scheme, and, vice-versa, each multiple kernel learning scheme defines a new norm.

The problem of kernel learning is a priori quite unrelated with parsimony. It emerges as a consequence of a convexity property of the

so-called "kernel trick", which we now describe. Consider a learning problem with $f(\boldsymbol{w}) = \psi(\boldsymbol{X}\boldsymbol{w})$. As seen before, this corresponds to linear predictions of the form $\boldsymbol{X}\boldsymbol{w} = (\boldsymbol{w}^\top \boldsymbol{x}_i)_{i=1,\dots,n}$. Assume that this learning problem is this time regularized by the square of the norm Ω (as shown in Section 1.2, this does not change the regularization properties), so the we have the following optimization problem:

$$\min_{\boldsymbol{w}\in\mathbb{R}^p} f(\boldsymbol{w}) + \frac{\lambda}{2}\Omega(\boldsymbol{w})^2. \tag{1.18}$$

As in Equation (1.12) we can introduce the linear constraint

$$\min_{\boldsymbol{u}\in\mathbb{R}^n,\boldsymbol{w}\in\mathbb{R}^p} \psi(\boldsymbol{u}) + \frac{\lambda}{2}\Omega(\boldsymbol{w})^2, \quad \text{s.t.} \quad \boldsymbol{u} = \boldsymbol{X}\boldsymbol{w}, \tag{1.19}$$

and reformulate the problem as the saddle point problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^n,\boldsymbol{w}\in\mathbb{R}^p} \max_{\alpha\in\mathbb{R}^n} \psi(\boldsymbol{u}) + \frac{\lambda}{2}\Omega(\boldsymbol{w})^2 - \lambda\boldsymbol{\alpha}^\top(\boldsymbol{u} - \boldsymbol{X}\boldsymbol{w}). \tag{1.20}$$

Since the primal problem (1.19) is a convex problem with feasible linear constraints, it satisfies Slater's qualification conditions and the order of maximization and minimization can be exchanged:

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} \min_{\boldsymbol{u}\in\mathbb{R}^n,\boldsymbol{w}\in\mathbb{R}^p} (\psi(\boldsymbol{u}) - \lambda\boldsymbol{\alpha}^\top\boldsymbol{u}) + \lambda\left(\frac{1}{2}\Omega(\boldsymbol{w})^2 + \boldsymbol{\alpha}^\top\boldsymbol{X}\boldsymbol{w}\right). \tag{1.21}$$

Now, the minimization in $\boldsymbol{u}$ and $\boldsymbol{w}$ can be performed independently. One property of norms is that the Fenchel conjugate of $\boldsymbol{w} \mapsto \frac{1}{2}\Omega(\boldsymbol{w})^2$ is $\boldsymbol{\kappa} \mapsto \frac{1}{2}\Omega^*(\boldsymbol{\kappa})^2$; this can be easily verified by finding the vector $\boldsymbol{w}$ achieving equality in the sequence of inequalities $\boldsymbol{\kappa}^\top\boldsymbol{w} \leq \Omega(\boldsymbol{w})\,\Omega^*(\boldsymbol{\kappa}) \leq \frac{1}{2}\big[\Omega(\boldsymbol{w})^2 + \Omega^*(\boldsymbol{\kappa})^2\big]$. As a consequence, the dual optimization problem is

$$\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\psi^*(\lambda\boldsymbol{\alpha}) - \frac{\lambda}{2}\Omega^*(\boldsymbol{X}^\top\boldsymbol{\alpha})^2. \tag{1.22}$$

If Ω is the Euclidean norm (i.e., the ℓ_2-norm) then the previous problem is simply

$$G(\boldsymbol{K}) := \max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\psi^*(\lambda\boldsymbol{\alpha}) - \frac{\lambda}{2}\boldsymbol{\alpha}^\top\boldsymbol{K}\boldsymbol{\alpha} \quad \text{with } \boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top. \tag{1.23}$$

Focussing on this last case, a few remarks are crucial:

(1) The dual problem depends on the design $\boldsymbol{X}$ only through the kernel matrix $\boldsymbol{K} = \boldsymbol{X}\boldsymbol{X}^\top \in \mathbb{R}^{n\times n}$.

(2) G is a *convex* function of $\boldsymbol{K}$ (as a maximum of linear functions).
(3) The solutions $\boldsymbol{w}^\star$ and $\boldsymbol{\alpha}^\star$ to the primal and dual problems satisfy $\boldsymbol{w}^\star = \boldsymbol{X}^\top \boldsymbol{\alpha}^\star = \sum_{i=1}^n \boldsymbol{\alpha}_i^\star \boldsymbol{x}_i$.
(4) The exact same duality result applies for the generalization to $\boldsymbol{w}, \boldsymbol{x}_i \in \mathcal{H}$ for $\mathcal{H}$ a Hilbert space.

The first remark suggests a way to solve learning problems that are non-linear in the inputs $\boldsymbol{x}_i$: in particular consider a non-linear mapping ϕ which maps $\boldsymbol{x}_i$ to a high-dimensional $\phi(\boldsymbol{x}_i) \in \mathcal{H}$ with $\mathcal{H} = \mathbb{R}^d$ for $d \gg p$ or possibly an infinite dimensional Hilbert space. Then consider the problem (1.18) with now $f(\boldsymbol{w}) = \psi\big((\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle)_{i=1,\dots,n}\big)$, which is typically of the form of an empirical risk $f(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^n \ell(y^{(i)}, \langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle)$. It becomes high-dimensional to solve in the primal, while it is simply solved in the dual by choosing a kernel matrix with entries $\boldsymbol{K}_{i,j} = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_j)\rangle$, which is advantageous as soon as $n^2 \leq d$; this is the so-called "kernel trick" (see more details in [122, 126]).

In particular, if we consider functions $h \in \mathcal{H}$ where $\mathcal{H}$ is a reproducing kernel Hilbert space (RKHS) with reproducing kernel K then

$$\min_{h \in \mathcal{H}} \psi\big((h(\boldsymbol{x}_i))_{i=1,\dots,n}\big) + \frac{\lambda}{2}\|h\|_{\mathcal{H}}^2 \tag{1.24}$$

is solved by solving Equation (1.23) with $\boldsymbol{K}_{i,j} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$. When applied to the mapping $\phi : \boldsymbol{x} \mapsto K(\boldsymbol{x}, \cdot)$, the third remark above yields a specific version of the representer theorem of Kimmeldorf and Wahba [74][12] stating that $h^\star(\cdot) = \sum_{i=1}^n \boldsymbol{\alpha}_i^\star K(\boldsymbol{x}_i, \cdot)$. In this case, the predictions may be written equivalently as $h(\boldsymbol{x}_i)$ or $\langle \boldsymbol{w}, \phi(\boldsymbol{x}_i)\rangle$, $i = 1, \dots, n$.

As shown in [77], the fact that G is a convex function of $\boldsymbol{K}$ suggests the possibility of optimizing the objective with respect to the choice of the kernel itself by solving a problem of the form $\min_{\boldsymbol{K} \in \mathcal{K}} G(\boldsymbol{K})$ where $\mathcal{K}$ is a convex set of kernel matrices.

In particular, given a finite set of kernel functions $(K_i)_{1 \leq i \leq p}$ it is natural to consider to find the best *linear* combination of kernels, which

[12] Note that this provides a proof of the representer theorem for *convex* losses only and that the parameters $\boldsymbol{\alpha}$ are obtained through a dual *maximization* problem.

requires to add a positive definiteness constraint on the kernel, leading to a semi-definite program [77]:

$$\min_{\boldsymbol{\eta}\in\mathbb{R}^p} G\left(\sum_{i=1}^p \boldsymbol{\eta}_i \boldsymbol{K}_i\right), \quad \text{s.t.} \quad \sum_{i=1}^p \eta_i \boldsymbol{K}_i \succeq 0, \quad \text{tr}\left(\sum_{i=1}^p \boldsymbol{\eta}_i \boldsymbol{K}_i\right) \leq 1. \tag{1.25}$$

Assuming that the kernels have equal trace, the two constraints of the previous program are avoided by considering convex combinations of kernels, which leads to a quadratically constrained quadratic program (QCQP) [78]:

$$\min_{\boldsymbol{\eta}\in\mathbb{R}^p_+} G\left(\sum_{i=1}^p \boldsymbol{\eta}_i \boldsymbol{K}_i\right), \text{ s.t. } \sum_{i=1}^p \boldsymbol{\eta}_i = 1. \tag{1.26}$$

We now present a reformulation of Equation (1.26) using sparsity-inducing norms (see [7, 13, 110] for more details).

1.5.1 From ℓ_1/ℓ_2-Regularization to MKL

As we presented it above, MKL arises from optimizing the objective of a learning problem w.r.t. to a convex combination of kernels, in the context of plain ℓ_2- or Hilbert norm regularization, which seems a priori unrelated to sparsity. We will show in this section that, in fact, the primal problem corresponding exactly to MKL (i.e., Equation 1.26) is an ℓ_1/ℓ_2-regularized problem (with the ℓ_1/ℓ_2-norm defined in Equation (1.2)), in the sense that its dual is the MKL problem for the set of kernels associated with each of the groups of variables. The proof to establish the relation between the two relies on the variational formulation presented in Section 1.4.2.

We indeed have, assuming that $\mathcal{G}$ is a partition of $\{1,\ldots,p\}$, with $|\mathcal{G}| = m$, and $\triangle_m$ denoting the simplex in $\mathbb{R}^m$,

$$\begin{aligned}\min_{\boldsymbol{w}\in\mathbb{R}^p} \psi(\boldsymbol{X}\boldsymbol{w}) + \frac{\lambda}{2}\left(\sum_{g\in\mathcal{G}} \|\boldsymbol{w}_g\|_2\right)^2 \\ = \min_{\boldsymbol{w}\in\mathbb{R}^p, \boldsymbol{\eta}\in\triangle_m} \psi(\boldsymbol{X}\boldsymbol{w}) + \frac{\lambda}{2}\sum_{g\in\mathcal{G}} \frac{\|\boldsymbol{w}_g\|_2^2}{\boldsymbol{\eta}_g}\end{aligned}$$

$$
\begin{aligned}
&= \min_{\widetilde{\boldsymbol{w}} \in \mathbb{R}^p, \boldsymbol{\eta} \in \triangle_m} \psi\left(\sum_{g \in \mathcal{G}} \boldsymbol{\eta}_g^{1/2} \boldsymbol{X}_g \widetilde{\boldsymbol{w}}_g\right) + \frac{\lambda}{2} \sum_{g \in \mathcal{G}} \|\widetilde{\boldsymbol{w}}_g\|_2^2 \\
&= \min_{\widetilde{\boldsymbol{w}} \in \mathbb{R}^p, \boldsymbol{\eta} \in \triangle_m} \psi(\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{w}}) + \frac{\lambda}{2} \|\widetilde{\boldsymbol{w}}\|_2^2, \text{ s.t. } \widetilde{\boldsymbol{X}} = [\eta_{g_1}^{1/2} \boldsymbol{X}_{g_1}, \ldots, \eta_{g_m}^{1/2} \boldsymbol{X}_{g_m}] \\
&= \min_{\boldsymbol{\eta} \in \triangle_m} \max_{\boldsymbol{\alpha} \in \mathbb{R}^n} -\psi^*(\lambda \boldsymbol{\alpha}) - \frac{\lambda}{2} \boldsymbol{\alpha}^\top \left(\sum_{g \in \mathcal{G}} \boldsymbol{\eta}_g \boldsymbol{K}_g\right) \boldsymbol{\alpha} \\
&= \min_{\boldsymbol{\eta} \in \triangle_m} G\left(\sum_{g \in \mathcal{G}} \boldsymbol{\eta}_g \boldsymbol{K}_g\right),
\end{aligned}
$$

where the third line results from the change of variable $\widetilde{\boldsymbol{w}}_g \boldsymbol{\eta}_g^{1/2} = \boldsymbol{w}_g$, and the last step from the definition of G in Equation (1.23).

Note that ℓ_1-regularization corresponds to the special case where groups are singletons and where $\boldsymbol{K}_i = \boldsymbol{x}_i \boldsymbol{x}_i^\top$ is a rank-one kernel matrix. In other words, MKL with rank-one kernel matrices (i.e., feature spaces of dimension one) is equivalent to ℓ_1-regularization (and thus simpler algorithms can be brought to bear in this situation).

We have shown that learning convex combinations of kernels through Equation (1.26) turns out to be equivalent to an ℓ_1/ℓ_2-norm penalized problems. In other words, learning a linear combination $\sum_{i=1}^m \boldsymbol{\eta}_i \boldsymbol{K}_i$ of kernel matrices, subject to $\boldsymbol{\eta}$ belonging to the simplex $\triangle_m$ is equivalent to penalizing the empirical risk with an ℓ_1-norm applied to norms of predictors $\|\boldsymbol{w}_g\|_2$. This link between the ℓ_1-norm and the simplex may be extended to other norms, among others to the sub-quadratic norms introduced in Section 1.4.2.

1.5.2 Structured Multiple Kernel Learning

In the relation established between ℓ_1/ℓ_2-regularization and MKL in the previous section, the vector of weights $\boldsymbol{\eta}$ for the different kernels corresponded to the vector of optimal variational parameters defining the norm. A natural way to extend MKL is, instead of considering a convex combination of kernels, to consider a linear combination of the same kernels, but with positive weights satisfying a different set of constraints than the simplex constraints. Given the relation between

kernel weights and the variational form of a norm, we will be able to show that, for norms that have a variational form as in Lemma 1.8, we can generalize the correspondence between the ℓ_1/ℓ_2-norm and MKL to a correspondence between other structured norms and structured MKL schemes.

Using the same line of proof as in the previous section, and given an H-norm (or equivalently a subquadratic norm) Ω_H as defined in Definition 1.2, we have:

$$\begin{aligned}
&\min_{\boldsymbol{w}\in\mathbb{R}^p} \psi(\boldsymbol{X}\boldsymbol{w}) + \frac{\lambda}{2}\Omega_H(\boldsymbol{w})^2 \\
&\quad= \min_{\boldsymbol{w}\in\mathbb{R}^p,\boldsymbol{\eta}\in H} \psi(\boldsymbol{X}\boldsymbol{w}) + \frac{\lambda}{2}\sum_{i=1}^p \frac{\boldsymbol{w}_i^2}{\boldsymbol{\eta}_i} \\
&\quad= \min_{\widetilde{\boldsymbol{w}}\in\mathbb{R}^p,\boldsymbol{\eta}\in H} \psi\left(\sum_{i=1}^p \boldsymbol{\eta}_i^{1/2}\boldsymbol{X}_i\widetilde{\boldsymbol{w}}_i\right) + \frac{\lambda}{2}\sum_{i=1}^p \widetilde{\boldsymbol{w}}_i^2 \\
&\quad= \min_{\widetilde{\boldsymbol{w}}\in\mathbb{R}^p,\boldsymbol{\eta}\in H} \psi(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{w}}) + \frac{\lambda}{2}\|\widetilde{\boldsymbol{w}}\|_2^2, \text{ s.t. } \widetilde{\boldsymbol{X}} = [\boldsymbol{\eta}_1^{1/2}\boldsymbol{X}_1,\dots,\boldsymbol{\eta}_p^{1/2}\boldsymbol{X}_p] \\
&\quad= \min_{\boldsymbol{\eta}\in H}\max_{\boldsymbol{\alpha}\in\mathbb{R}^n} -\psi^*(\lambda\boldsymbol{\alpha}) - \frac{\lambda}{2}\boldsymbol{\alpha}^\top\left(\sum_{i=1}^p \boldsymbol{\eta}_i\boldsymbol{K}_i\right)\boldsymbol{\alpha} \\
&\quad= \min_{\boldsymbol{\eta}\in H} G\left(\sum_{i=1}^p \boldsymbol{\eta}_i\boldsymbol{K}_i\right). \qquad (1.27)
\end{aligned}$$

This results shows that the regularization with the norm Ω_H in the primal is equivalent to a multiple kernel learning formulation in which the kernel weights are constrained to belong to the convex set H, which defines Ω_H variationally. Note that we have assumed that $H \subset \mathbb{R}_+^p$, so that formulations such as (1.25), where positive semidefiniteness of $\sum_{i=1}^p \boldsymbol{\eta}_i\boldsymbol{K}_i$ has to be added as a constraint, are not included.

Given the relationship of MKL to the problem of learning a function in a reproducing kernel Hilbert space, the previous result suggests a natural extension of structured sparsity to the RKHS settings. Indeed let, $h = (h_1,\dots,h_p) \in \mathcal{B} := \mathcal{H}_1 \times \cdots \times \mathcal{H}_p$, where $\mathcal{H}_i$ are RKHSs. It is easy to verify that $\Lambda: h \mapsto \Omega_H\big((\|h_1\|_{\mathcal{H}_1},\dots,\|h_p\|_{\mathcal{H}_p})\big)$ is a convex function, using the variational formulation of Ω_H, and since it is also

non-negative definite and homogeneous, it is a norm.[13] Moreover, the learning problem obtained by summing the predictions from the different RKHSs, i.e.,

$$\min_{h\in\mathcal{B}} \psi((h_1(\boldsymbol{x}_i) + \cdots + h_p(\boldsymbol{x}_i))_{i=1,\ldots,n}) + \frac{\lambda}{2}\Omega_H((\|h_1\|_{\mathcal{H}_1},\ldots,\|h_p\|_{\mathcal{H}_p}))^2 \tag{1.28}$$

is equivalent, by the above derivation, to the MKL problem $\min_{\boldsymbol{\eta}\in H} G(\sum_{i=1}^p \boldsymbol{\eta}_i \boldsymbol{K}_i)$ with $[\boldsymbol{K}_i]_{j,j'} = K_i(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$ for K_i the reproducing kernel of $\mathcal{H}_i$. See Section 3.4 for more details.

This means that, for most of the structured sparsity-inducing norms that we have considered in Section 1.3, we may replace individual variables by whole Hilbert spaces. For example, tree-structured sparsity (and its extension to directed acyclic graphs) was explored in [9] where each node of the graph was an RKHS, with an application to nonlinear variable selection.

[13] As we show in Section 3.4, it is actually sufficient to assume that Ω is monotonic for Λ to be a norm.

References

[1] J. Abernethy, F. Bach, T. Evgeniou, and J.-P. Vert, "A new approach to collaborative filtering: Operator estimation with spectral regularization," *Journal of Machine Learning Research*, vol. 10, pp. 803–826, 2009.

[2] J. Aflalo, A. Ben-Tal, C. Bhattacharyya, J. S. Nath, and S. Raman, "Variable sparsity kernel learning," *Journal of Machine Learning Research*, vol. 12, pp. 565–592, 2011.

[3] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on Signal Processing*, vol. 54, no. 11, pp. 4311–4322, 2006.

[4] Y. Amit, M. Fink, N. Srebro, and S. Ullman, "Uncovering shared structures in multiclass classification," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2007.

[5] C. Archambeau and F. Bach, "Sparse probabilistic projections," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[6] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[7] F. Bach, "Consistency of the group lasso and multiple kernel learning," *Journal of Machine Learning Research*, vol. 9, pp. 1179–1225, 2008.

[8] F. Bach, "Consistency of trace norm minimization," *Journal of Machine Learning Research*, vol. 9, pp. 1019–1048, 2008.

[9] F. Bach, "Exploring large feature spaces with hierarchical multiple kernel learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[10] F. Bach, "Structured sparsity-inducing norms through submodular functions," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[11] F. Bach, "Shaping level sets with submodular functions," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.
[12] F. Bach, R. Jenatton, J. Mairal, and G. Obozinski, "Convex optimization with sparsity-inducing norms," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. J. Wright, eds.), MIT Press, 2011.
[13] F. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2004.
[14] F. Bach, J. Mairal, and J. Ponce, "Convex sparse matrix factorizations," *Preprint arXiv:0812.1869v1*, 2008.
[15] R. G. Baraniuk, V. Cevher, M. Duarte, and C. Hegde, "Model-based compressive sensing," *IEEE Transactions on Information Theory*, vol. 56, no. 4, pp. 1982–2001, 2010.
[16] F. L. Bauer, J. Stoer, and C. Witzgall, "Absolute and monotonic norms," *Numerische Mathematik*, vol. 3, no. 1, pp. 257–264, 1961.
[17] A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
[18] D. P. Bertsekas, *Nonlinear Programming*. Athena Scientific, 2nd ed., 1999.
[19] P. Bickel, Y. Ritov, and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector," *Annals of Statistics*, vol. 37, no. 4, pp. 1705–1732, 2009.
[20] J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization: Theory and Examples*. Springer-Verlag, 2006.
[21] L. Bottou, "Online algorithms and stochastic approximations," *Online Learning and Neural Networks*, vol. 5, 1998.
[22] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.
[23] L. Bottou and Y. LeCun, "Large scale online learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2004.
[24] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, "Distributed optimization and statistical learning via the alternating direction method of multipliers," *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–124, 2011.
[25] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.
[26] D. M. Bradley and J. A. Bagnell, "Convex coding," in *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence (UAI)*, 2009.
[27] P. Brucker, "An $O(n)$ algorithm for quadratic knapsack problems," *Operations Research Letters*, vol. 3, no. 3, pp. 163–166, 1984.
[28] C. Burges, "Dimension reduction: A guided tour," *Machine Learning*, vol. 2, no. 4, pp. 275–365, 2009.
[29] J. F. Cai, E. J. Candès, and Z. Shen, "A singular value thresholding algorithm for matrix completion," *SIAM Journal on Optimization*, vol. 20, pp. 1956–1982, 2010.
[30] E. J. Candès, M. Wakin, and S. Boyd, "Enhancing sparsity by reweighted L1 minimization," *Journal of Fourier Analysis and Applications*, vol. 14, no. 5, pp. 877–905, 2008.

[31] F. Caron and A. Doucet, "Sparse Bayesian nonparametric regression," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[32] V. Cehver, M. Duarte, C. Hedge, and R. G. Baraniuk, "Sparse signal recovery using markov random fields," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[33] A. Chambolle, "Total variation minimization and a class of binary MRF models," in *Proceedings of the fifth International Workshop on Energy Minimization Methods in Computer Vision and Pattern Recognition*, 2005.

[34] A. Chambolle and J. Darbon, "On total variation minimization and surface evolution using parametric maximum flows," *International Journal of Computer Vision*, vol. 84, no. 3, pp. 288–307, 2009.

[35] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The convex geometry of linear inverse problems," *Preprint arXiv:1012.0621*, 2010.

[36] G. H. G. Chen and R. T. Rockafellar, "Convergence rates in forward-backward splitting," *SIAM Journal on Optimization*, vol. 7, no. 2, pp. 421–444, 1997.

[37] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM Journal on Scientific Computing*, vol. 20, pp. 33–61, 1999.

[38] P. L. Combettes and J.-C. Pesquet, *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*. Chapter Proximal Splitting Methods in Signal Processing. Springer-Verlag, 2011.

[39] P. L. Combettes and V. R. Wajs, "Signal recovery by proximal forward-backward splitting," *SIAM Multiscale Modeling and Simulation*, vol. 4, no. 4, pp. 1168–1200, 2006.

[40] S. F. Cotter, J. Adler, B. Rao, and K. Kreutz-Delgado, "Forward sequential algorithms for best basis selection," in *IEEE Proceedings of Vision Image and Signal Processing*, pp. 235–244, 1999.

[41] I. Daubechies, R. DeVore, M. Fornasier, and C. S. Güntürk, "Iteratively reweighted least squares minimization for sparse recovery," *Communications on Pure and Applied Mathematics*, vol. 63, no. 1, pp. 1–38, 2010.

[42] D. L. Donoho and I. M. Johnstone, "Adapting to unknown smoothness via wavelet shrinkage," *Journal of the American Statistical Association*, vol. 90, no. 432, pp. 1200–1224, 1995.

[43] J. Duchi, E. Hazan, and Y. Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, pp. 2121–2159, 2011.

[44] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least angle regression," *Annals of Statistics*, vol. 32, no. 2, pp. 407–499, 2004.

[45] M. Elad and M. Aharon, "Image denoising via sparse and redundant representations over learned dictionaries," *IEEE Transactions on Image Processing*, vol. 15, no. 12, pp. 3736–3745, 2006.

[46] K. Engan, S. O. Aase, and H. Husoy et al., "Method of optimal directions for frame design," in *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 1999.

[47] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American Statistical Association*, vol. 96, no. 456, pp. 1348–1360, 2001.

[48] M. Fazel, H. Hindi, and S. Boyd, "A rank minimization heuristic with application to minimum order system approximation," in *Proceedings of the American Control Conference*, vol. 6, pp. 4734–4739, 2001.

[49] J. Friedman, T. Hastie, and R. Tibshirani, "A note on the group lasso and a sparse group lasso," *Preprint arXiv:1001:0736v1*, 2010.

[50] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[51] W. J. Fu, "Penalized regressions: The bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397–416, 1998.

[52] G. Gasso, A. Rakotomamonjy, and S. Canu, "Recovering sparse signals with non-convex penalties and DC programming," *IEEE Transactions on Signal Processing*, vol. 57, no. 12, pp. 4686–4698, 2009.

[53] A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.

[54] R. Glowinski and P. Le Tallec, *Augmented Lagrangian and Operator-splitting Methods in Nonlinear Mechanics*. Society for Industrial Mathematics, 1989.

[55] Y. Grandvalet and S. Canu, "Outcomes of the equivalence of adaptive ridge with least absolute shrinkage," in *Advances in Neural Information Processing Systems (NIPS)*, 1999.

[56] R. Gribonval, V. Cevher, and M. E. Davies, "Compressible distributions for high-dimensional statistics," *preprint arXiv:1102.1249v2*, 2011.

[57] Z. Harchaoui, "Méthodes à Noyaux pour la Détection," PhD thesis, Télécom ParisTech, 2008.

[58] Z. Harchaoui and C. Lévy-Leduc, "Catching change-points with Lasso," in *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[59] K. K. Herrity, A. C. Gilbert, and J. A. Tropp, "Sparse approximation via iterative thresholding," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing, (ICASSP)*, 2006.

[60] C. Hu, J. Kwok, and W. Pan, "Accelerated gradient methods for stochastic optimization and online learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[61] J. Huang and T. Zhang, "The benefit of group sparsity," *Annals of Statistics*, vol. 38, no. 4, pp. 1978–2004, 2010.

[62] J. Huang, Z. Zhang, and D. Metaxas, "Learning with structured sparsity," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[63] H. Ishwaran and J. S. Rao, "Spike and slab variable selection: frequentist and Bayesian strategies," *Annals of Statistics*, vol. 33, no. 2, pp. 730–773, 2005.

[64] L. Jacob, G. Obozinski, and J.-P. Vert, "Group Lasso with overlaps and graph Lasso," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[65] R. Jenatton, "Structured sparsity-inducing norms: Statistical and algorithmic properties with applications to neuroimaging," PhD thesis, Ecole Normale Supérieure de Cachan, 2012.

[66] R. Jenatton, J.-Y. Audibert, and F. Bach, "Structured variable selection with sparsity-inducing norms," *Journal of Machine Learning Research*, vol. 12, pp. 2777–2824, 2011.

[67] R. Jenatton, A. Gramfort, V. Michel, G. Obozinski, F. Bach, and B. Thirion, "Multi-scale mining of fMRI data with hierarchical structured sparsity," in *International Workshop on Pattern Recognition in Neuroimaging (PRNI)*, 2011.

[68] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for sparse hierarchical dictionary learning," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[69] R. Jenatton, J. Mairal, G. Obozinski, and F. Bach, "Proximal methods for hierarchical sparse coding," *Journal of Machine Learning Research*, vol. 12, pp. 2297–2334, 2011.

[70] R. Jenatton, G. Obozinski, and F. Bach, "Structured sparse principal component analysis," in *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[71] S. C. Johnson, "Hierarchical clustering schemes," *Psychometrika*, vol. 32, no. 3, pp. 241–254, 1967.

[72] K. Kavukcuoglu, M. A. Ranzato, R. Fergus, and Y. LeCun, "Learning invariant features through topographic filter maps," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[73] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2010.

[74] G. S. Kimeldorf and G. Wahba, "Some results on Tchebycheffian spline functions," *Journal of Mathematical Analysis and Applications*, vol. 33, pp. 82–95, 1971.

[75] K. Koh, S. J. Kim, and S. Boyd, "An interior-point method for large-scale l1-regularized logistic regression," *Journal of Machine Learning Research*, vol. 8, no. 8, pp. 1519–1555, 2007.

[76] B. Krishnapuram, L. Carin, M. A. T. Figueiredo, and A. J. Hartemink, "Sparse multinomial logistic regression: Fast algorithms and generalization bounds," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 6, pp. 957–968, 2005.

[77] G. R. G. Lanckriet, N. Cristianini, L. El Ghaoui, P. Bartlett, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[78] G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, "A statistical framework for genomic data fusion," *Bioinformatics*, vol. 20, pp. 2626–2635, 2004.

[79] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[80] A. Lefèvre, F. Bach, and C. Févotte, "Itakura-Saito nonnegative matrix factorization with group sparsity," in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2011.

[81] P. L. Lions and B. Mercier, "Splitting algorithms for the sum of two nonlinear operators," *SIAM Journal on Numerical Analysis*, vol. 16, no. 6, pp. 964–979, 1979.

[82] H. Liu, M. Palatucci, and J. Zhang, "Blockwise coordinate descent procedures for the multi-task lasso, with applications to neural semantic basis discovery," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[83] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer, "Taking advantage of sparsity in multi-task learning," *Preprint arXiv:0903.1468*, 2009.

[84] N. Maculan and G. Galdino de Paula, "A linear-time median-finding algorithm for projecting a vector on the simplex of $\mathbb{R}^n$," *Operations Research Letters*, vol. 8, no. 4, pp. 219–222, 1989.

[85] J. Mairal, "Sparse coding for machine learning, image processing and computer vision," PhD thesis, Ecole Normale Supérieure de Cachan, `http://tel.archives-ouvertes.fr/tel-00595312`, 2010.

[86] J. Mairal, F. Bach, J. Ponce, and G. Sapiro, "Online learning for matrix factorization and sparse coding," *Journal of Machine Learning Research*, vol. 11, pp. 19–60, 2010.

[87] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Network flow algorithms for structured sparsity," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[88] J. Mairal, R. Jenatton, G. Obozinski, and F. Bach, "Convex and network flow optimization for structured sparsity," *Journal of Machine Learning Research*, vol. 12, pp. 2681–2720, 2011.

[89] S. Mallat and Z. Zhang, "Matching pursuit in a time-frequency dictionary," *IEEE Transactions on Signal Processing*, vol. 41, no. 12, pp. 3397–3415, 1993.

[90] H. Markowitz, "Portfolio selection," *Journal of Finance*, vol. 7, no. 1, pp. 77–91, 1952.

[91] B. Martinet, "Régularisation d'inéquations variationnelles par approximations successives," *Revue franaise d'informatique et de recherche opérationnelle, série rouge*, 1970.

[92] A. F. T. Martins, N. A. Smith, P. M. Q. Aguiar, and M. A. T. Figueiredo, "Structured sparsity in structured prediction," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2011.

[93] C. A. Micchelli, J. M. Morales, and M. Pontil, "Regularizers for structured sparsity," *Preprint arXiv:1010.0556v2*, 2011.

[94] J. J. Moreau, "Fonctions convexes duales et points proximaux dans un espace hilbertien," *Comptes-Rendus de l'Académie des Sciences de Paris, Série A, Mathématiques*, vol. 255, pp. 2897–2899, 1962.

[95] S. Mosci, L. Rosasco, M. Santoro, A. Verri, and S. Villa, "Solving structured sparsity regularization with proximal methods," *Machine Learning and Knowledge Discovery in Databases*, pp. 418–433, 2010.

[96] B. K. Natarajan, "Sparse approximate solutions to linear systems," *SIAM Journal on Computing*, vol. 24, pp. 227–234, 1995.

[97] R. M. Neal, *Bayesian Learning for Neural Networks*, vol. 118. Springer Verlag, 1996.

[98] D. Needell and J. Tropp, "CoSaMP: Iterative signal recovery from incomplete and inaccurate samples," *Applied and Computational Harmonic Analysis*, vol. 26, no. 3, pp. 301–321, 2009.

[99] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu, "A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[100] Y. Nesterov, *Introductory Lectures on Convex Optimization: A Basic Course*. Kluwer Academic Publishers, 2004.

[101] Y. Nesterov, "Smooth minimization of non-smooth functions," *Mathematical Programming*, vol. 103, no. 1, pp. 127–152, 2005.

[102] Y. Nesterov, "Gradient methods for minimizing composite objective function," Technical Report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Discussion paper, 2007.

[103] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems," Technical Report, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain, Discussion paper, 2010.

[104] J. Nocedal and S. J. Wright, *Numerical Optimization*. Springer Verlag, 2nd ed., 2006.

[105] G. Obozinski, L. Jacob, and J.-P. Vert, "Group Lasso with overlaps: the Latent Group Lasso approach," *preprint HAL — inria-00628498*, 2011.

[106] G. Obozinski, B. Taskar, and M. I. Jordan, "Joint covariate selection and joint subspace selection for multiple classification problems," *Statistics and Computing*, vol. 20, no. 2, pp. 231–252, 2009.

[107] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, pp. 607–609, 1996.

[108] M. R. Osborne, B. Presnell, and B. A. Turlach, "On the Lasso and its dual," *Journal of Computational and Graphical Statistics*, vol. 9, no. 2, pp. 319–337, 2000.

[109] M. Pontil, A. Argyriou, and T. Evgeniou, "Multi-task feature learning," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[110] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL," *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[111] N. S. Rao, R. D. Nowak, S. J. Wright, and N. G. Kingsbury, "Convex approaches to model wavelet sparsity patterns," in *International Conference on Image Processing (ICIP)*, 2011.

[112] F. Rapaport, E. Barillot, and J.-P. Vert, "Classification of arrayCGH data using fused SVM," *Bioinformatics*, vol. 24, no. 13, pp. 375–382, 2008.

[113] P. Ravikumar, J. Lafferty, H. Liu, and L. Wasserman, "Sparse additive models," *Journal of the Royal Statistical Society. Series B, statistical methodology*, vol. 71, pp. 1009–1030, 2009.

[114] K. Ritter, "Ein verfahren zur lösung parameterabhängiger, nichtlinearer maximum-probleme," *Mathematical Methods of Operations Research*, vol. 6, no. 4, pp. 149–166, 1962.

[115] R. T. Rockafellar, *Convex Analysis*. Princeton University Press, 1997.

[116] V. Roth and B. Fischer, "The Group-Lasso for generalized linear models: uniqueness of solutions and efficient algorithms," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

[117] L. I. Rudin, S. Osher, and E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D: Nonlinear Phenomena*, vol. 60, no. 1-4, pp. 259–268, 1992.

[118] M. Schmidt, G. Fung, and R. Rosales, "Fast optimization methods for L1 regularization: A comparative study and two new approaches," in *Proceedings of the European Conference on Machine Learning (ECML)*, 2007.

[119] M. Schmidt, D. Kim, and S. Sra, "Projected Newton-type methods in machine learning," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. J. Wright, eds.), MIT Press, 2011.

[120] M. Schmidt, N. Le Roux, and F. Bach, "Convergence rates of inexact proximal-gradient methods for convex optimization," in *Advances in Neural Information Processing Systems (NIPS)*, 2011.

[121] M. Schmidt and K. Murphy, "Convex structure learning in log-linear models: Beyond pairwise potentials," in *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[122] B. Schölkopf and A. J. Smola, *Learning with Kernels*. MIT Press, 2001.

[123] M. W. Seeger, "Bayesian inference and optimal design for the sparse linear model," *Journal of Machine Learning Research*, vol. 9, pp. 759–813, 2008.

[124] S. Shalev-Shwartz and A. Tewari, "Stochastic methods for ℓ_1-regularized loss minimization," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

[125] A. Shapiro, D. Dentcheva, and A. P. Ruszczyński, *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial Mathematics, 2009.

[126] J. Shawe-Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[127] S. K. Shevade and S. S. Keerthi, "A simple and efficient algorithm for gene selection using sparse logistic regression," *Bioinformatics*, vol. 19, no. 17, pp. 2246–2253, 2003. Oxford Univ Press.

[128] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning," *Journal of Machine Learning Research*, vol. 7, pp. 1531–1565, 2006.

[129] P. Sprechmann, I. Ramirez, G. Sapiro, and Y. C. Eldar, "C-HiLasso: A collaborative hierarchical sparse modeling framework," *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4183–4198, 2011.

[130] N. Srebro, J. D. M. Rennie, and T. S. Jaakkola, "Maximum-margin matrix factorization," in *Advances in Neural Information Processing Systems (NIPS)*, 2005.

[131] T. Suzuki and R. Tomioka, "SpicyMKL: A fast algorithm for multiple kernel learning with thousands of kernels," *Machine Learning*, vol. 85, pp. 77–108, 2011.

[132] M. Szafranski, Y. Grandvalet, and P. Morizet-Mahoudeaux, "Hierarchical penalization," in *Advances in Neural Information Processing Systems (NIPS)*, 2007.

[133] R. Tibshirani, "Regression shrinkage and selection via the Lasso," *Journal of the Royal Statistical Society Series B*, vol. 58, no. 1, pp. 267–288, 1996.

[134] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused Lasso," *Journal of the Royal Statistical Society Series B*, vol. 67, no. 1, pp. 91–108, 2005.

[135] R. Tomioka, T. Suzuki, and M. Sugiyama, "Augmented Lagrangian methods for learning, selecting and combining features," in *Optimization for Machine Learning*, (S. Sra, S. Nowozin, and S. J. Wright, eds.), MIT Press, 2011.

[136] J. A. Tropp, "Greed is good: Algorithmic results for sparse approximation," *IEEE Transactions on Signal Processing*, vol. 50, no. 10, pp. 2231–2242, 2004.

[137] J. A. Tropp, A. C. Gilbert, and M. J. Strauss, "Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit," *Signal Processing, Special Issue "Sparse Approximations in Signal and Image Processing"*, vol. 86, pp. 572–588, 2006.

[138] P. Tseng, "Applications of a splitting algorithm to decomposition in convex programming and variational inequalities," *SIAM Journal on Control and Optimization*, vol. 29, pp. 119–138, 1991.

[139] P. Tseng, "On accelerated proximal gradient methods for convex-concave optimization," *submitted to SIAM Journal on Optimization*, 2008.

[140] P. Tseng and S. Yun, "A coordinate gradient descent method for nonsmooth separable minimization," *Mathematical Programming*, vol. 117, no. 1, pp. 387–423, 2009.

[141] B. A. Turlach, W. N. Venables, and S. J. Wright, "Simultaneous variable selection," *Technometrics*, vol. 47, no. 3, pp. 349–363, 2005.

[142] E. Van den Berg, M. Schmidt, M. P. Friedlander, and K. Murphy, "Group sparsity via linear-time projections," Technical report, University of British Columbia, Technical Report number TR-2008-09, 2008.

[143] G. Varoquaux, R. Jenatton, A. Gramfort, G. Obozinski, B. Thirion, and F. Bach, "Sparse structured dictionary learning for brain resting-state activity modeling," in *NIPS Workshop on Practical Applications of Sparse Modeling: Open Issues and New Directions*, 2010.

[144] J.-P. Vert and K. Bleakley, "Fast detection of multiple change-points shared by many signals using group LARS," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[145] M. J. Wainwright, "Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming," *IEEE Transactions on Information Theory*, vol. 55, no. 5, pp. 2183–2202, 2009.

[146] S. Weisberg, *Applied Linear Regression*. Wiley, 1980.

[147] D. P. Wipf and S. Nagarajan, "A new view of automatic relevance determination," *Advances in Neural Information Processing Systems (NIPS)*, 2008.

[148] D. P. Wipf and B. D. Rao, "Sparse bayesian learning for basis selection," vol. 52, no. 8, pp. 2153–2164, 2004.

[149] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[150] S. J. Wright, "Accelerated block-coordinate relaxation for regularized optimization," Technical report, Technical report, University of Wisconsin-Madison, 2010.

[151] S. J. Wright, R. D. Nowak, and M. A. T. Figueiredo, "Sparse reconstruction by separable approximation," *IEEE Transactions on Signal Processing*, vol. 57, no. 7, pp. 2479–2493, 2009.

[152] T. T. Wu and K. Lange, "Coordinate descent algorithms for lasso penalized regression," *Annals of Applied Statistics*, vol. 2, no. 1, pp. 224–244, 2008.

[153] L. Xiao, "Dual averaging methods for regularized stochastic learning and online optimization," *Journal of Machine Learning Research*, vol. 9, pp. 2543–2596, 2010.

[154] H. Xu, C. Caramanis, and S. Sanghavi, "Robust PCA via outlier pursuit," *Preprint arXiv:1010.4237*, 2010.

[155] G. X. Yuan, K. W. Chang, C. J. Hsieh, and C. J. Lin, "A comparison of optimization methods for large-scale l1-regularized linear classification," Technical Report, Department of Computer Science, National University of Taiwan, 2010.

[156] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B*, vol. 68, pp. 49–67, 2006.

[157] P. Zhao, G. Rocha, and B. Yu, "The composite absolute penalties family for grouped and hierarchical variable selection," *Annals of Statistics*, vol. 37, no. 6A, pp. 3468–3497, 2009.

[158] P. Zhao and B. Yu, "On model selection consistency of Lasso," *Journal of Machine Learning Research*, vol. 7, pp. 2541–2563, 2006.

[159] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B*, vol. 67, no. 2, pp. 301–320, 2005.