Interactions between rank and sparsity in penalized estimation, and detection of structured objects

Pierre-André Savalle

Ecole Centrale Paris

Thèse de Doctorat

Spécialité: Mathématiques

Présentée par
Pierre-André Savalle

Interactions entre Rang et Parcimonie en Estimation Pénalisée, et Détection d'Objets Structurés

Soutenue le 21/10/2014 devant le jury composé de:

M. Alexandre	**D'ASPREMONT**	Ecole Normale Supérieure	Rapporteur
M. Arnak	**DALALYAN**	CREST, ENSAE, Uni. Paris-Est	Rapporteur
M. Gilles	**FAY**	Ecole Centrale Paris	Directeur
M. Charles-Albert	**LEHALLE**	Capital Fund Management	Examinateur
M. Guillaume	**OBOZINSKI**	Uni. Paris-Est, Ecole des Ponts	Examinateur
M. Nicolas	**VAYATIS**	ENS Cachan	Directeur

Thèse #: 2014ECAP0051

Contents

1 Introduction **15**
- 1.1 Contributions: Interactions Between Rank and Sparsity . . . 17
  - 1.1.1 Estimation of Sparse and Low Rank Matrices . . . 17
  - 1.1.2 Convex Localized Multiple Kernel Learning . . . 19
- 1.2 Contributions: Detection of Structured Objects . . . 20
  - 1.2.1 Detection of Correlations with Adaptive Sensing . . . 21
  - 1.2.2 Detection of Objects with High-dimensional CNN Features . . . 23

I Interactions Between Rank and Sparsity in Penalized Estimation **27**

2 Penalized Matrix Estimation **29**
- 2.1 Introduction . . . 30
  - 2.1.1 Linear Models: Generalization and Estimation . . . 30
  - 2.1.2 Occam's Razor and Minimum Description Length . . . 33
  - 2.1.3 Priors and Penalized Estimation . . . 34
  - 2.1.4 Penalized Matrix Estimation . . . 35
  - 2.1.5 Penalized or Constrained? . . . 37
- 2.2 Sparsity . . . 38
  - 2.2.1 The ℓ_0-norm . . . 38
  - 2.2.2 The ℓ_1-norm . . . 39
  - 2.2.3 Elastic-net and Variations . . . 41
  - 2.2.4 Other Regularizers . . . 42
- 2.3 Rank and Latent Factor Models . . . 43
  - 2.3.1 The Rank . . . 44
  - 2.3.2 The Trace Norm . . . 44
  - 2.3.3 Nuclear and Atomic Norms . . . 47
  - 2.3.4 The Max Norm . . . 49
  - 2.3.5 Other Heuristics and Open Problems . . . 50
- 2.4 Measuring Quality of Regularizers and Theoretical Results . . . 51
  - 2.4.1 Estimation: Exact and Robust Recovery . . . 52
  - 2.4.2 High-Dimensional Convex Sets and Gaussian Width . . . 53
  - 2.4.3 Optimality Conditions for Penalized Estimation . . . 54
  - 2.4.4 Kinematic Formula and Statistical Dimension . . . 57
  - 2.4.5 Examples . . . 58
  - 2.4.6 Estimation with Non-gaussian Designs . . . 60

3 Estimation of Sparse and Low Rank Matrices **63**
- 3.1 Introduction . . . 64
  - 3.1.1 Model . . . 65
  - 3.1.2 Main Examples . . . 65

3.1.3 Outline . . . 65
3.1.4 Notation . . . 66
3.2 Oracle Inequality . . . 66
3.3 Generalization Error in Link Prediction . . . 67
3.4 Algorithms . . . 68
3.4.1 Proximal Operators . . . 68
3.4.2 Generalized Forward-backward Splitting . . . 69
3.4.3 Incremental Proximal Descent . . . 69
3.4.4 Positive Semi-definite Constraint . . . 70
3.5 Recovering Clusters . . . 70
3.6 Numerical Experiments . . . 71
3.6.1 Synthetic Data . . . 71
3.6.2 Real Data Sets . . . 71
3.7 Discussion . . . 73
3.7.1 Other Loss Functions . . . 73
3.7.2 Optimization . . . 74
3.7.3 Geometry . . . 74
3.7.4 Factorization Methods . . . 75
3.8 Proofs . . . 78
3.8.1 Sketch of Proof of Proposition 13 . . . 78
3.8.2 Proof of Proposition 16 . . . 79

4 Convex Localized Multiple Kernel Learning 81
4.1 Introduction . . . 82
4.1.1 Linear MKL . . . 82
4.1.2 Hinge Loss Aggregation . . . 83
4.1.3 Related Work . . . 84
4.1.4 Outline . . . 86
4.1.5 Notation . . . 86
4.2 Generalized Hinge Losses . . . 86
4.2.1 Representer Theorem . . . 88
4.2.2 Universal Consistency . . . 89
4.3 ℓ_p-norm Aggregation of Hinge Losses . . . 91
4.3.1 Dual Problem . . . 91
4.3.2 Regularization . . . 93
4.3.3 Link Functions . . . 94
4.4 Experiments . . . 94
4.4.1 UCI Datasets . . . 95
4.4.2 Image Classification . . . 96
4.5 Discussion . . . 97
4.6 Proofs . . . 98
4.6.1 Proof of Lemma 2 . . . 98
4.6.2 Proof of Lemma 3 . . . 98
4.6.3 Proof of Lemma 4 . . . 99
4.6.4 Proof of Lemma 5 . . . 99
4.6.5 Proof of Theorem 3 . . . 100

II Detection of Structured Objects **101**

5 Detection of Correlations with Adaptive Sensing **103**
5.1 Introduction . . . 104
5.1.1 Model . . . 104
5.1.2 Adaptive vs. Non-adaptive Sensing and Testing . . . 105
5.1.3 Uniform Sensing and Testing . . . 107
5.1.4 Related Work . . . 108
5.1.5 Outline . . . 109
5.1.6 Notation . . . 109
5.2 Lower Bounds . . . 110
5.3 Adaptive Tests . . . 113
5.3.1 Sequential Thresholding . . . 113
5.3.2 The Case of k-intervals . . . 115
5.3.3 The Case of k-sets: Randomized Subsampling . . . 118
5.4 Unnormalized Correlation Model . . . 120
5.4.1 Model and Extensions of Previous Results . . . 120
5.4.2 The Case of k-sets . . . 120
5.5 Discussion . . . 122
5.6 Proofs . . . 124
5.6.1 Inequalities and KL Divergences . . . 124
5.6.2 Proof of Bound on KL Divergence . . . 125
5.6.3 Proof of Proposition 18 . . . 125
5.6.4 Proof of Proposition 19 . . . 126
5.7 Appendix: Extensions to Unnormalized Model . . . 127
5.7.1 Uniform (non-adaptive) Lower Bound . . . 127
5.7.2 Uniform (non-adaptive) Upper Bound . . . 127
5.7.3 KL Divergences . . . 127

6 Detection of Objects with CNN Features **131**
6.1 Introduction . . . 132
6.2 Detection with DPMs . . . 133
6.2.1 Detection Task . . . 133
6.2.2 Deformable Part Models . . . 133
6.3 Integrating Convolutional Features into DPMs . . . 135
6.3.1 The Alexnet Network Structure . . . 135
6.3.2 Prior Work . . . 136
6.3.3 Using CNN Layer 5 Features in DPMs . . . 137
6.4 Results on Pascal VOC 2007 . . . 138

Bibliography **141**

Résumé

Cette thèse est organisée en deux parties indépendantes. La première partie s'intéresse à l'estimation convexe de matrice en prenant en compte à la fois la parcimonie et le rang. Dans le contexte de graphes avec une structure de communautés, on suppose souvent que la matrice d'adjacence sous-jacente est diagonale par blocs dans une base appropriée. Cependant, de tels graphes possèdent généralement une matrice d'adjacente qui est aussi parcimonieuse, ce qui suggère que combiner parcimonie et rang puisse permettre de modéliser ce type d'objet de manière plus fine. Nous proposons et étudions ainsi une pénalité convexe pour promouvoir parcimonie et rang faible simultanément. Même si l'hypothèse de rang faible permet de diminuer le sur-apprentissage en diminuant la capacité d'un modèle matriciel, il peut être souhaitable lorsque suffisamment de données sont disponibles de ne pas introduire une telle hypothèse. Nous étudions un exemple dans le contexte multiple kernel learning localisé, où nous proposons une famille de méthodes á vaste-marge convexes et accompagnées d'une analyse théorique. La deuxième partie de cette thèse s'intéresse à des problèmes de détection d'objets ou de signaux structurés. Dans un premier temps, nous considérons un problème de test statistique, pour des modèles où l'alternative correspond à des capteurs émettant des signaux corrélés. Contrairement à la littérature traditionnelle, nous considérons des procédures de test séquentielles, et nous établissons que de telles procédures permettent de détecter des corrélations significativement plus faibles que les méthodes traditionnelles. Dans un second temps, nous considérons le problème de localiser des objets dans des images. En s'appuyant sur de récents résultats en apprentissage de représentation pour des problèmes similaires, nous intégrons des features de grande dimension issues de réseaux de neurones convolutionnels dans les modèles déformables traditionnellement utilisés pour ce type de problème. Nous démontrons expérimentalement que ce type d'approche permet de diminuer significativement le taux d'erreur de ces modèles.

Abstract

This thesis is organized in two independent parts. The first part focused on convex matrix estimation problems, where both rank and sparsity are taken into account simultaneously. In the context of graphs with community structures, a common assumption is that the underlying adjacency matrices are block-diagonal in an appropriate basis. However, these types of graphs are usually far from complete, and their adjacency representations are thus also inherently sparse. This suggests that combining the sparse hypothesis and the low rank hypothesis may allow to more accurately model such objects. To this end, we propose and analyze a convex penalty to promote low rank and high sparsity simultaneously. Although the low rank hypothesis allows to reduce over-fitting by decreasing the modeling capacity of a matrix model, the opposite may be desirable when enough data is available. We study such an example in the context of localized multiple kernel learning, which extends multiple kernel learning by allowing each of the kernels to select different support vectors. In this framework, multiple kernel learning corresponds to a rank one estimator, while higher-rank estimators have been observed to increase generalization performance. We propose a novel family of large-margin methods for this problem that, unlike previous methods, are both convex and theoretically grounded. The second part of the thesis is about detection of objects or signals which exhibit combinatorial structures, and we present two such problems. First, we consider detection in the statistical hypothesis testing sense, in models where anomalous signals correspond to correlated values at different sensors. In most existing work, detection procedures are provided with a full sample of all the sensors. However, the experimenter may have the capacity to make targeted measurements in an on-line and adaptive manner, and we investigate such adaptive sensing procedures. Finally, we consider the task of identifying and localizing objects in images. This is an important problem in computer vision, where hand-crafted features are usually used. Following recent successes in learning ad-hoc representations for similar problems, we integrate the method of deformable part models with high-dimensional features from convolutional neural networks, and shows that this significantly decreases the error rates of existing part-based models.

Acknowledgements

First, I would like to thank Gilles and Nicolas, for never loosing faith that this thesis would lead somewhere, providing me with a much coveted mix of safety and freedom. I hope that this is only the beginning of many more adventures together. My deepest thanks to Alexandre d'Aspremont and Arnak Dalalyan for doing me the honor of agreeing to review this thesis. I am also grateful to Charles-Albert Lehalle and Guillaume Obozinski, for participating in my thesis committee, and for taking interest in my work. I am more than humbled to have you all on my committee.

I also would like to thank everyone I have had the opportunity to work with these last three years: all my thanks to Emile, for so many heated discussions and good moments; to Gábor, for welcoming me so warmly in Barcelona, and showing me how blissful riding the roller coaster of research can be; to Iasonas, who swore never again to stay up late on deadlines, and yet is still trying out *the next idea* around the clock; to Rui, who knows more about parisian rooftops than I might ever do. My thoughts and my gratitude also go to everyone that has supported me in these last years; to everyone I've crossed path with at CMLA; to Rémy, who never backs away from a fight with a taco or a mathematical conjecture; to everyone at MAS for scores of unforgettable moments, and most particularly to Alex, Benjamin, Benoit, Blandine, Charlotte, Marion, Ludovic, Gautier; to Annie, Micheline, Sylvie and Virginie; to Nikos, for consistently reminding me of upcoming deadlines, and consistently believing in me; to Benoit, who has more ideas by the minute than any reasonable person has by the day, and who has been a constant support throughout these three years; to Jérémy, my sparing partner, who is feared by wavelet-sparse signals and bacon toppings across the entire universe; to all my friends and my family, and, most importantly, to Hélène and my parents.

Finally, I want to dedicate this work to Ben Taskar, who is missed by all more than my words can clumsily express.

Notation

General

$[d]$	$\{1,\dots,d\}$	
$\operatorname{diam}(X)$	Diameter of set X in Euclidean norm	p. 54
Π_C	Projection operator onto the convex set C	p. 57
$\operatorname{diag}(M)$	Vector of diagonal elements of matrix M	
$X \circ Y$	Hadamard product of X and Y	p. 36
M_+	Componentwise positive part of matrix M	
$\operatorname{Tr}(M)$	Trace of matrix M	p. 45
$M \succeq 0$	Matrix M is positive semi-definite	p. 45
$\ker M$	Kernel of matrix M	p. 52
$\mathbf{1}$	All ones vector	p. 30
$\mathbf{1}[C]$	Zero-one indicator function of condition C	p. 32
δ_X	Barrier indicator function of set X	p. 40
$\operatorname{conv} X$	Convex hull of set X	p. 42
$\operatorname{cone} X$	Conical hull of set X	p. 54
$f^\star$	Convex conjugate of function f	p. 40
$f^{\star\star}$	Convex biconjugate of function f	p. 40
KL	Kullback-Leibler divergence	p. 110

Matrix Norms and Balls

$\|\cdot\|^\star$	Dual norm to $\|\cdot\|$	p. 47
$\|M\|_0 = \operatorname{card}\{M_{i,j} \neq 0\}$	ℓ_0-norm of matrix M	
$\|M\|_p = \left[\sum_{i,j} \|M_{i,j}\|^p\right]^{1/p}$	ℓ_p-norm of matrix M	
$\|M\|_1 = \sum_{i,j} \|M_{i,j}\|$	ℓ_1-norm of matrix M	
$\|M\|_F = \sqrt{\sum_{i,j} M_{i,j}^2}$	Frobenius norm of matrix M	
$\|M\|_\infty = \max \|M_{i,j}\|$	ℓ_∞-norm of matrix M	
$\|M\|_{\text{op}} = \sup_{x\,:\,\|x\|_2=1} \|Mx\|_2$	Operator norm of matrix M	
$\|M\|_*$	Trace norm of matrix M	p. 45
$\|M\|_{\max}$	Max norm of matrix M	p. 49
$\mathbf{B}_\Omega$	Unit ball of norm Ω	p. 34
$\mathbf{S}_\Omega$	Unit sphere of norm Ω	p. 34
$w(C)$	Gaussian width of set C	p. 53
$\Delta(C)$	Statistical dimension of set C	p. 57
$\mathcal{N}(X, \|.\|, \varepsilon)$	ε-covering number of X with respect to norm $\|.\|$	p. 89
$H(X, \|.\|, \varepsilon)$	Metric entropy (logarithm of covering number)	p. 89

1
Introduction

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not 'Eureka!' (I found it!) but 'That's funny ...'

— Isaac Asimov

The cognitive process in the brain can be simplified down to two processes around which thinking is structured: induction and deduction. In induction, we build new mental models from observations, while in deduction, we use these mental models to make predictions and take actions. These two phases are continuously alternating, and although deduction is easy, induction is fundamentally harder and ill-defined. At a high-level, many methods in machine learning can be understood as such continuously alternating phases of induction and deduction. This is particularly visible in methods which aim at learning a feature representation from the data, such as in dictionary learning, where induction corresponds to the estimation of codewords, and deduction to the coding problem. Convex methods correspond to a rather simplified modeling environment where induction and deduction are actually done jointly through some well-grounded optimization algorithm. Although the cognitive process is a highly non-convex process, simple mental models can be useful to get some initial insight on a problem or a concept at a moderate cost. Similarly, convex methods have proven dramatically useful to approach many statistical modeling and learning problems, and we believe that advancing the state of knowledge in what can be effectively handled with this type of methods is of utmost interest. This does not mean, however, that non-convex methods should be avoided, and many such methods have recently proven very effective at achieving state of the art results in a variety of problems. Instead, we believe that the classical debate of which of convex or non-convex methods are best is actually moot, and that these two classes correspond to *largely different trade-offs*. If one were to draw a cartoon picture, convex methods are usually associated with lower computational and modeling efforts, and can often be used to quickly gain significant insight on a given problem. Non-convex methods are generally significantly heavier to deploy and require an arguably more hands-on expertise, but have repeatedly proven effective at going further than convex methods in many learning and pattern recognition tasks.

This thesis is organized in two independent parts. This chapter summarizes our results and contributions. The first part of the thesis is on convex estimation problems, with an emphasis on combining classical hypotheses usually handled in isolation for matrix

estimation. We consider how objects such as graphs with community structure, covariance matrices or mixtures of kernel machines can be modeled in a convex framework, and we focus in particular on models involving rank and sparsity hypotheses. The second part of the thesis is on two examples of detection problems, where objects or signals to be detected have some type of combinatorial structure. This departs largely from convex models, as we consider both information-theoretically optimal procedures and numerically successful methods which are highly non-convex.

1.1 Contributions: Interactions Between Rank and Sparsity

The first part of this thesis is focused on penalized estimation problems for regression and classification. More specifically, we consider convex problems for estimating matrices. A key element which differentiate this problem from standard high dimensional vector estimation is that different structural assumptions can be formulated in this context: although sparsity hypotheses can be transposed from the vector case, genuinely different hypotheses can be formulated, such as the low rank hypothesis.

The concepts of sparsity and of low rank have been central in statistics and machine learning in the last decade, and have been at the source of numerous successes. At the core of the sparsity hypothesis lies the idea that the data may be well modeled by a limited number of features or variables, and that performing *variable selection* may increase the predictive performance (Mallat, 1999; Bühlmann and Van De Geer, 2011). The low rank hypothesis is at the source of a variety of models usually referred to as *latent factor models*, which are also widely recognized as effective in practice, such as in clustering (Shahnaz et al., 2006), recommender systems (Koren, 2008), or blind source separation (Cichocki et al., 2009).

Chapter 2 is dedicated to reviewing the ideas underlying sparsity and of low rank models. We also briefly retrace the history of penalized estimation, and present general geometric tools which can be used to study the estimation performance of a large array of convex penalties from a theoretical point of view. In Chapter 3 and Chapter 4, we consider two matrix estimation problems, where *both* rank and sparsity are taken into account simultaneously, albeit in different ways. In the following, we present a preview of the results from these two works.

1.1.1 Estimation of Sparse and Low Rank Matrices

In recent years, the notion of sparsity for vectors has been transposed into the concept of low rank matrices, and this latter hypothesis has opened up the way to numerous achievements (Srebro, 2004; Cai et al., 2008). In Chapter 3, we argue that being low rank is not only an equivalent of sparsity for matrices but also that low rank and sparsity can actually be seen as two orthogonal concepts. The underlying structure we have in mind is that of a block diagonal matrix. This situation occurs for instance in covariance matrix estimation in the case of groups of highly correlated variables or when clustering social graphs.

Consider the adjacency matrix of a graph with community structure, as illustrated on the leftmost panel of Figure 1.1: this is characterized by fully (or densely) connected blocks, or *clusters*. This can be translated into the hypothesis that the matrix is *low rank*. In many graph problems, a central assumption is that we are observing a noisy or partial version of a graph with this kind of structure, as pictured in the second leftmost panel of Figure 1.1. This low rank assumption is used, for instance, in matrix completion to predict unobserved movie ratings. Methods based on low rank approximations or low rank inducing penalties such as the trace norm usually yield dense matrices as estimators. This is shown on the third panel of Figure 1.1: even after thresholding of small magnitude entries for display purposes, the support is still quite dense. Although this may be fine for predicting ratings, this is not adapted for graph adjacency matrices, which are usually sparse. This is the case, in particular, for most communication networks, for social networks, or for biological interaction networks. By using a sparsity inducing penalty in addition of the low rank

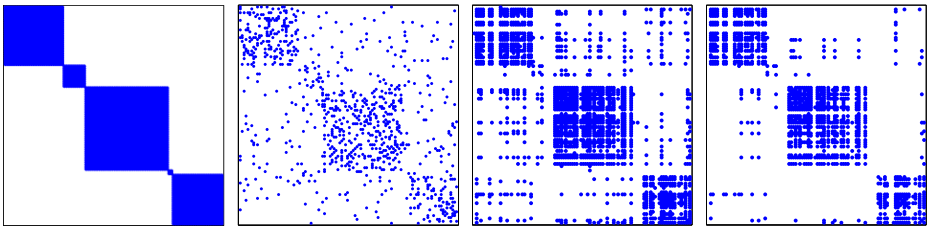

Figure 1.1: Adjacency matrix with community structure, sparse regularization, low rank regularization, and sparse low rank regularization

inducing penalty, we obtain the support shown on the rightmost panel of Figure 1.1, which is a much better recovery of the original graph structure.

The problem of leveraging both types of structures at the same time is largely different from *demixing* or *decomposition* problems (Amelunxen et al., 2013) such as Robust PCA (Candès et al., 2011; Chandrasekaran et al., 2011), where the objective is to recover a sparse matrix $S \in \mathbb{R}^{d_1 \times d_2}$ and a low rank matrix $L \in \mathbb{R}^{d_1 \times d_2}$ from the knowledge of their sum $X = L + S$ only. For this type of problem, objective functions usually take the form of an infimal convolution (Rockafellar, 1997) of penalties (Agarwal et al., 2012). On the other hand, we consider here a *single* matrix that is *simultaneously* sparse and low rank.

Contributions. We propose a novel convex penalty to encourage solutions that achieve a tradeoff between low rank and high sparsity. The penalty is based on a linear combination of the classical surrogates, the matrix ℓ_1-norm, and the trace norm : for a matrix $W \in \mathbb{R}^{d \times d}$, the penalty is

$$\gamma \|W\|_1 + (1-\gamma)\|W\|_*.$$

We derive oracle inequalities for penalized estimation using this penalty.

Proposition 1. *Let $W^\star \in \mathbb{R}^{d \times d}$ and $A = W^\star + \epsilon$ with $\epsilon \in \mathbb{R}^{d \times d}$ having i.i.d. entries with zero mean. Assume for some $\alpha \in [0,1]$ that $\tau \geq 2\alpha\|\epsilon\|_{op}$ and $\gamma \geq 2(1-\alpha)\|\epsilon\|_\infty$. Let*

$$\widehat{W} = \operatorname*{arg\,min}_{W \in \mathbb{R}^{d \times d}} \Big[\|W - A\|_F^2 + \gamma \|W\|_1 + \tau \|W\|_* \Big].$$

Then,

$$\|\widehat{W} - W^\star\|_F^2 \leq \inf_{W \in \mathbb{R}^{d \times d}} \Big[\|W - W^\star\|_F^2 + 2\gamma\|W\|_1 + 2\tau\|W\|_* \Big],$$

$$\|\widehat{W} - W^\star\|_F^2 \leq \Big[2\gamma\|W^\star\|_1 + 2\tau\|W^\star\|_* \Big] \wedge \left[\gamma\sqrt{\|W^\star\|_0} + \tau\sqrt{\operatorname{rank}(W^\star)}\,\frac{\sqrt{2}+1}{2} \right]^2.$$

This is extended in Proposition 13 to constrained optimization (for instance over the cone of semi-definite positive matrices), and generalizes previous sharp bounds for Lasso or trace norm regression (Koltchinskii et al., 2011a). Combined with proper choices of regularization parameters that we discuss, this allows to control the prediction error. Using proximal methods, we demonstrate the benefits of combining these two hypotheses both on synthetic data, and on real data such as protein interaction data and social networks. This chapter is an extended version of a paper with Emile Richard and Nicolas Vayatis which has appeared in the proceedings of ICML (Richard et al., 2012).

Future directions. We propose a convex relaxation to the intersection of the two classical manifolds that are sparse matrices and low rank matrices. However, it may well be that there are some more interesting joint measure of rank and sparsity to relax when modeling the type of data we are interested in. We discuss possible directions based on matrix factorizations and atomic norms in Section 3.7. More generally, the following problem appears of interest: *how can multiple priors on a model be combined to estimate the model with fewer observations, and for what prior structures is this a significant improvement over using a single prior?* For instance, Kamal and Vandergheynst (2013) consider combining rank and sparsity, but in different bases. An interesting direction to tackle this problem could be *learning* to combine such priors, as done with kernels and Hilbert space metrics in multiple kernel learning.

1.1.2 Convex Localized Multiple Kernel Learning

The low rank hypothesis allows to decrease the modeling capacity of a matrix model, which may be helpful in a high-dimensional setting to avoid over-fitting. In other settings, however, the opposite may be desirable: more complex models may allow to obtain better generalization performance if enough data is available. We consider in Chapter 4 such an examples, in the context of localized multiple kernel learning. Kernel-based methods such as SVMs are very effective tools for classification and regression. In addition to good empirical performance in a wide range of situations, these method and backed by a strong theoretical background (Steinwart and Christmann, 2008), as well as mature algorithms (Platt, 1999) and implementations (Chang and Lin, 2011). The kernel is traditionally either selected from generic parametric off-the-shelf kernels (e.g., Gaussian or polynomial) using some sort of cross-validation, or hand-crafted by domain specific experts through an expensive empirical process. Multiple kernel learning (MKL) has been proposed to alleviate part of this expensive model selection problem (Lanckriet et al., 2002; Bach et al., 2004; Gönen and Alpaydın, 2011). In MKL, the SVM formulation is modified to jointly learn a classifier and linear combination weights for a set $K_1, \dots, K_M$ of kernels. This results in classifiers of the form

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i \underbrace{\left[\sum_{m=1}^{M} s_m K_m(x_i, x) \right]}_{\text{kernel is a mixture of kernels}},$$

where $\alpha \in \mathbb{R}^N$ and $s_1, \dots, s_M \in \mathbb{R}_+$ are to be both learned at the same time. In the wake of the success of MKL, there has been recent interest in combining kernels in a localized or data-dependent way (Bi et al., 2004; Gönen and Alpaydin, 2008; Cao et al., 2009; Gehler and Nowozin, 2009), as opposed to more traditional linear combinations where all the kernels must agree on a common set of support vectors. By allowing each kernel to select separate and possibly different support vectors, these may be more relevant with respect to the information encoded by the kernels. These approaches are referred to as either *localized*, or *data-dependent* multiple kernel learning.

In the chapter, we focus on fully nonparametric localized MKL classifiers of the form

$$f(x) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_i \alpha_{i,m} K_m(x_i, x),$$

where $\alpha \in \mathbb{R}^{N \times M}$ is a matrix. While the solution to the linear MKL problem can be written in this form with $\mathrm{rank}(\alpha) = 1$, we are interested in *higher rank solutions* which exhibit localization, in the sense that the relative weights of the kernels vary depending on the support vectors. Informally, this corresponds to a kernel combination that is different in different regions of the feature space. Previous methods are either non-convex, are not large margin methods, or consider only parametric models for α.

Contributions. We propose a family of large-margin methods that are both convex and theoretically grounded for combining kernels in a data-dependent manner, based on aggregating hinge losses over each of the kernels. For $p \in [1, \infty]$, we consider large margin programs of the form

$$\min_{\omega=(\omega_1,\ldots,\omega_M)\in H} \left[\frac{1}{2} \sum_{m=1}^{M} \|\omega_m\|_2^2 + C \sum_{i=1}^{N} \left[\sum_{m=1}^{M} (1 - y_i \langle \omega_1, \phi_m(x_i) \rangle)_+^p \right]^{1/p} \right],$$

where $\phi_1, \ldots, \phi_M$ are the feature mappings associated to the kernels. This allows through p to adjust the amount of coupling between the kernels. When $p = \infty$, kernels are the most tightly coupled, while when $p = 1$, the method amounts to averaging the decisions of independently trained SVMs. We show that classifiers defined from these programs (and actually more general aggregations) are universally consistent, and we consider the question of whether *intermediate levels of coupling* can be beneficial in practice.

We evaluate these methods on real data, including both UCI datasets, and image classification tasks from computer vision. Our experimental validation includes multiple methods which have not been previously compared, although they address a similar question. Our experimental validation shows that $p = 1$ (i.e., averaging the decisions of independent SVMs) is superior to any other value of p, but also to standard MKL and previously introduced methods for localized MKL. In addition of being by far the simplest and cheapest to implement and run, this yields the best (or close to the best) classification accuracy in all of our benchmarks. Similarly to how the simple average kernel often achieves performances comparable to that of MKL (Gehler and Nowozin, 2009), this suggests that straightforward methods may achieve close to state-of-the-art accuracies for localized MKL as well. This chapter is joint work with Antoine Poliakov, and has been submitted.

1.2 Contributions: Detection of Structured Objects

Detection problems are a wide and pervasive class of problem, where the high-level goal is to detect expected or unexpected patterns from observations and data. This type of problem differs significantly from matrix estimation problems as presented in the first part of this thesis, in that the models and structures involved are usually radically different. In the second part of this thesis, we consider two types of detection problems involving structured target objects, from statistical hypothesis testing and computer vision.

In the first problem, we consider detection in the statistical hypothesis testing sense: we want to known *whether* there is some anomalous signal, but we do not seek to identify precisely where. We characterize the minimax risk in different models where anomalous signals correspond to *correlated* values at different sensors. In the second problem, we consider the more applied task of identifying different types of real-world objects in images. The *detection* terminology has a different meaning here than in statistical testing:

the objective is to precisely localizing where in the image these object are. We integrate the method of *deformable part models* with high-dimensional features from *convolutional neural networks*, and shows that this significantly decreases the error rate of DPMs.

1.2.1 Detection of Correlations with Adaptive Sensing

The first detection problem that we consider is a statistical hypothesis testing problem. Given multiple observations from a Gaussian multivariate distribution, we want to test whether the corresponding covariance matrix is diagonal against non-diagonal alternatives. More precisely, we consider a testing problem over a Gaussian vector $U \in \mathbb{R}^n$, where under the alternative hypothesis, there exists an unknown subset S of $\{1,\dots,n\}$ such that the corresponding components are positively correlated with strength $\rho \in]0,1[$, while the others are independent.

This can be interpreted as an anomaly detection problem: picture a spatially arranged array of sensors. In the normal regime, signals at each of the sensors consists only of zero-mean uncorrelated Gaussian noise. In the anomalous regime, some of the sensors instead have correlated signals. This has to be distinguished from many classical anomaly detection problems where one assume that the anomaly is characterized by an elevated signal mean at some of the sensors. In the model that we consider, anomalies cannot be detected by looking at sensors in isolation, but only when considering correlations between multiple sensors.

The models that we consider are very similar to the *rank one spiked covariance model*, which has been associated in recent years to sparse PCA (Johnstone and Lu, 2009; Berthet and Rigollet, 2013; Cai et al., 2013). We focus on detection of positive correlations (Arias-Castro et al., 2012, 2014), and we consider the *sparse* regime where only a relatively small number of the n components are correlated, if any. The subset S can be any subset of size of a known size k (which we refer to as the k-sets problem), or may have additional structure known to the experimenter. For instance, it can consists of k contiguous coordinates in $\{1,\dots,n\}$, in which case one expects that detection will be easier due to this extra information. This last setting is referred to as the k-intervals problem, and can be generalized for instance to rectangles $\{i_0,\dots,i_0+k_1-1\}\times\{j_0,\dots,j_0+k_2-1\}$ with $k_1k_2 = k$, when the n coordinates are arranged spatially on a two dimensional grid $\{1,\dots,n_1\}\times\{1,\dots,n_2\}$ with $n_1n_2 = n$.

In the litterature (Hero and Rajaratnam, 2012; Arias-Castro et al., 2014, 2012; Berthet and Rigollet, 2013; Cai et al., 2013), this problem or related problems have been analyzed under *uniform sensing*, where i.i.d. draws $U_1,\dots,U_m \in \mathbb{R}^n$ of the Gaussian vector are available. Our approach deviates from this in that we consider an *adaptive sensing* or *sequential experimental design* setting. More precisely, data is collected in a sequential and adaptive way, where data collected at earlier stages informs the collection of data in future stages. In particular, the experimenter may choose to acquire only a subset of the coordinates from the Gaussian vector. In the previous sensor array illustration, a cost may be associated to obtaining a measurement from a sensor, and the experimenter may choose to activate only specific subsets of sensors. This is illustrated in Figure 1.2. Here, coordinates $\{1,\dots,n\}$ are laid out according to a grid. The set of correlated coordinates is a convex shape, and these are shown in red. These coordinates form a clique in the graph of correlations, and this is shown through light red edges. At every step, the experimenter selects coordinates to be sensed, and these are shown circled. At the first step, the exper-

imenter samples all the coordinates, while at the two subsequent steps, the experimenter reduced the amount of coordinates sampled. This illustrates an important point: we only

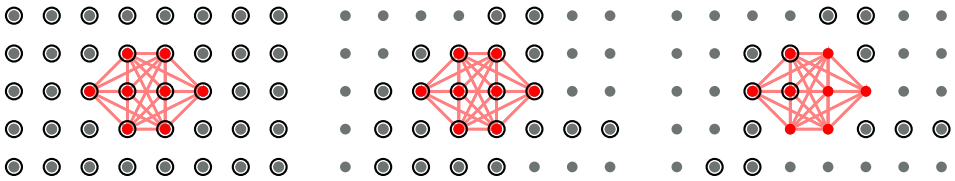

Figure 1.2: Adaptive sensing over a two dimensional grid of sensors

consider the testing problem of finding out *whether* there are correlated coordinates, not the problem of estimating which. As a consequence, it may be fine for the experimenter to discard some coordinates from S. Indeed, all that is needed is that, under the alternative hypothesis, we may detect at least two correlated components with some certainty, as this is sufficient to reject the null hypothesis.

Adaptive sensing has been studied in the context of other detection and estimation problems, such as in detection of a shift in the mean of a Gaussian vector (Castro, 2012; Haupt et al., 2009), in compressed sensing (Arias-Castro et al., 2013; Haupt et al., 2012; Castro, 2012), in experimental design, optimization with Gaussian processes (Srinivas et al., 2010), and in active learning (Chen and Krause, 2013). Adaptive sensing procedures are quite flexible, as the data collection procedure can be "steered" to ensure most collected data provides important information. As a consequence, procedures based on adaptive sensing are often associated with better detection or estimation performances than those based on non-adaptive sensing with a similar measurement budget.

In non-adaptive sensing, all the decisions regarding the collection of data must be taken before any observations are made, which generalizes slightly the setting of uniform sensing. As already mentioned, uniform sensing corresponds to the case where m full vectors are observed, corresponding to a total of $M = mn$ coordinate measures. This problem has been thoroughly studied in (Arias-Castro et al., 2014). To allow for easier comparison with the special case of uniform sensing, we will use adaptive sensing with a budget of M coordinate measurements, and we ultimately express our results in terms of m, the equivalent number of full vector measurements.

We are interested in the *high-dimensional* setting, where the ambient dimension n is high. All quantities such as the correlation coefficient ρ, the correlated set size $k = o(n)$, and the number of vector measurements m will thus be allowed to depend on n, and to go to infinity simultaneously, albeit possibly at different rates. We seek to identify the range of parameters in which it is possible to construct adaptive tests whose minimax risks converge to zero.

Contributions. We show in Chapter 5 that adaptive sensing procedures can significantly outperform the best non-adaptive tests both for k-intervals and k-sets. In the following, we provide a preview of our results for the case where $\rho k \to 0$ (that is, the total amount of correlation is asymptotically vanishing), although we also provide results for the case where $\rho k \to \infty$. The constants are omitted.

For k-intervals, our necessary and sufficient conditions for asymptotic detection are almost matching. In particular, the number of measurements m necessary and sufficient to

ensure that the risk approaches zero has essentially no dependence on the signal dimension n:

$$\textbf{Necessary:} \quad \rho k\sqrt{m} \to \infty, \qquad \textbf{Sufficient:} \quad \rho k\sqrt{m} \geq \sqrt{\log\log\frac{n}{k}}.$$

This is in stark contrast with the non-adaptive sensing results, where it is sufficient (and almost necessary) for m to grow logarithmically with n according to

$$\rho k\sqrt{m} \geq \sqrt{\log\frac{n}{k}}.$$

This type of dependence that is almost independent of the dimension has been observed before in the context of adaptive sensing: due to the ability to sequentially adapt the experimental design, the experimenter may abstract himself almost completely from the original problem dimension.

For k-sets, we obtain a sufficient condition that still depend logarithmically in n, but which improves nonetheless upon uniform sensing in some regimes:

$$\textbf{Sufficient:} \quad \rho\sqrt{km} \geq \sqrt{\log\frac{n}{k}} \quad \text{and} \quad \rho km \geq \log\frac{n}{k}.$$

This should be compared to uniform sensing, where it is sufficient (and, again, almost necessary) when $k = o(\sqrt{n})$ that

$$\rho\sqrt{km} \geq \sqrt{\log n} \quad \text{and} \quad \rho m \geq \log n.$$

In addition to this, in a slightly different model (a rank-one spiked covariance model), we obtain a tighter sufficient condition for detection of k-sets, that is nearly independent of the dimension n, and also improves significantly over non-adaptive sensing:

$$\textbf{Sufficient:} \quad \rho\sqrt{km} \geq \log\log\frac{n}{k}.$$

This chapter is joint work with Rui Castro and Gábor Lugosi (Castro et al., 2013).

Future directions. For k-sets, we obtain the same lower bound as for k-intervals: for detection to be possible, it must hold that $\rho k\sqrt{m}$ goes to infinity. While this lower bound is tight for k-intervals, we do not know if this is the case for k-sets, as the dependence in k does not match that of our sufficient condition. This leaves open an important question: *does the structure of the correlated set help when using adaptive sensing*? This is the case for uniform sensing, and may appear reasonable for adaptive sensing as well. However, in a similar study on adaptive testing for *elevated means* (as opposed to correlations), many symmetric classes of correlated sets such as k-sets or k-intervals have been shown to *all* have minimax risk which converge to zero under the *same* necessary and sufficient conditions (Castro, 2012), such that structure does not help in this setting.

1.2.2 Detection of Objects with High-dimensional CNN Features

The second detection problem that we consider is a computer vision problem, where one is given images, and should detect and precisely localize objects of different types. This

is an important problem in computer vision, and has been the subject of a large body of work. Although significant progress has been made in recent years (Sermanet et al., 2014; Girshick et al., 2014), error rates are still significant, ranging from 30% to 60% depending on the object type, and averaging at 40% for the state of the art methods on the 20-classes PASCAL VOC 2007 dataset (Everingham et al., 2007, 2010b). The problem differs from the classification task of finding the class of a single dominant object in the image. Instead, in the task that we consider, multiples objects of identical or different types may be present, and their location must be predicted accurately.

Challenges for detection are multiple. The detection of a given object type requires to correctly learn to separate this object from other similar looking objects (and from background), but also to precisely pinpoint the position of the object in the image. Driven partly by the availability of larger datasets and partly by increasing industrial demand, the interest for detection of a large number of classes of objects has raised recently. Computational considerations pose additional challenges when working with such datasets.

A general approach for object detection is that of *sliding windows*: for all possible patches of a predefined size, we run a classifier to predict whether or not this corresponds to a given type of object. A natural approach consists in transforming each window into a feature vector, and training a binary black-box classifier to detect whether the window correspond to the object. However, objects in natural images are usually subject to a large variability in terms of orientation, scale, structure, illumination, or general appearance, and such monolithic classifiers over the window may not lead to the best performance. Instead, Bag-of-Words (BOW) models represent an image or object as the unstructured collection of all its patches of a given size. This simplifying representation was originally used in natural language processing and information retrieval, but has been the subject of significative interest in the vision community. In this context, BoW models are also referred to as *bags of visual words* (Yang et al., 2007). Unlike with approaches based on monolithic classifiers, BOW models may allow to independently detect small distinctive features of objects, and can hence be more robust to variabilities in natural images. However, these models do not allow to model any kind of structure.

Part-based models have received a lot of interest due to their ability to handle variabilities similarly as with BOW, while allowing to model the spatial structure of the visual words. Consider the task of learning to detect persons: this can be broken up into the arguably easier tasks of learning to detect heads, feet, legs, torsos, and of learning in what spatial arrangement these parts usually appear in images. This is the main idea behind *deformable part models* (DPMs), and has been largely successful (Everingham et al., 2010b). DPMs are sliding window detectors that are *structured*. In practice, models do not enforce any kind of interpretability of the parts, unlike what we mentioned. Formally, part positions are treated as latent variables, and detections correspond to windows for which one can find a highly-scoring latent part configuration, as illustrated in Figure 1.3. On the 20-classes PASCAL VOC 2007 dataset, Felzenszwalb et al. (2010b) have achieved error rates of 70% using DPMs.

Like numerous methods in computer vision, DPMs are based on hand-crafted image features: in (Felzenszwalb et al., 2010b), Histograms of Gradient features (also referred to as HOG) are used. Numerous other features have been proposed, including SIFT (Lowe, 1999), Haar wavelet coefficients (Viola and Jones, 2001), Shape Contexts (Belongie et al., 2002) or Local Binary Patterns (Ojala et al., 2002). Recently, techniques aiming at learning the feature representation directly from the data have proved extremely effective in various

Figure 1.3: Part based models, sample detections from Felzenszwalb et al. (2010b): detection bounding box (red), latent parts (blue)

domains including computer vision. In particular, convolutional neural networks (CNNs) have achieved state of the art error rates on image classification tasks (Krizhevsky et al., 2012), which suggested that such techniques could be of interest for detection as well.

This was recently confirmed by the R-CNN (Girshick et al., 2014) method for detection, where promising regions are obtained from segmentation considerations and warped to a fixed size window, which is then used as input to a CNN as in classification tasks. R-CNN achieves a dramatic improvement with respect to the state of art (Fidler et al., 2013), with a mean error rate of about 40% over the 20 classes of the PASCAL VOC 2007 dataset. However, this type of method based on training monolithic classifiers must be contrasted with structured detectors such as DPMs. Indeed, reducing the problem to a classification problem for fixed size windows does not allow for as much modeling flexibility as in part-based models. For instance, part-based method are inherently structured, which render them extendable to much more general problems such as human pose estimation (Yang and Ramanan, 2011), facial expression recognition (Zhu and Ramanan, 2012b), or three-dimensional structure estimation (Kakadiaris et al., 2007). However, none of this is possible with the type of approach used in R-CNN.

Contributions. In Chapter 6, we show how to integrate features from CNNs in the framework of DPMs. This constitutes a challenge: compared to HOG features, this corresponds to an eight fold increase in the dimension (from 32 to 256), within a framework which is already quite computationally expensive. Due to computational efficiency consideration, we use features computed from convolutional layers only. This is unlike the features used in classification or in R-CNN, for which features are computed using extra fully-connected layers on top of the convolutional layers. We demonstrate an increase of up to +9.7% in mean average precision (mean AP) with respect to DPMs on the PASCAL VOC 2007 dataset. This chapter is joint work with Iasonas Kokkinos and Stavros Tsogkas.

Future directions. Although we are able to significantly improve with respect to HOG-based DPMs, the mean AP that we achieve is still below the performance of recent methods such as R-CNN when using features from fully-connected layers. However, using only features from convolutional layers, we still achieve a mean AP close to what R-CNN achieves based only on these same layers. This suggests that adding further nonlinearities on top of our framework may help.

Part I

Interactions Between Rank and Sparsity in Penalized Estimation

2

Penalized Matrix Estimation

Contents

2.1 Introduction . . . 30
2.1.1 Linear Models: Generalization and Estimation . . . 30
2.1.2 Occam's Razor and Minimum Description Length . . . 33
2.1.3 Priors and Penalized Estimation . . . 34
2.1.4 Penalized Matrix Estimation . . . 35
2.1.5 Penalized or Constrained? . . . 37
2.2 Sparsity . . . 38
2.2.1 The ℓ_0-norm . . . 38
2.2.2 The ℓ_1-norm . . . 39
2.2.3 Elastic-net and Variations . . . 41
2.2.4 Other Regularizers . . . 42
2.3 Rank and Latent Factor Models . . . 43
2.3.1 The Rank . . . 44
2.3.2 The Trace Norm . . . 44
2.3.3 Nuclear and Atomic Norms . . . 47
2.3.4 The Max Norm . . . 49
2.3.5 Other Heuristics and Open Problems . . . 50
2.4 Measuring Quality of Regularizers and Theoretical Results . . . 51
2.4.1 Estimation: Exact and Robust Recovery . . . 52
2.4.2 High-Dimensional Convex Sets and Gaussian Width . . . 53
2.4.3 Optimality Conditions for Penalized Estimation . . . 54
2.4.4 Kinematic Formula and Statistical Dimension . . . 57
2.4.5 Examples . . . 58
2.4.6 Estimation with Non-gaussian Designs . . . 60

2.1 Introduction

In this thesis, we consider two ubiquitous problems of statistical learning: classification and regression. In both settings, one is given points $x_1, \dots, x_N \in \mathbb{R}^d$ as well as associated target values (or *labels*) that we denote by $y_1, \dots, y_N \in \mathcal{Y}$. The target values are to be predicted from the points. The points are also referred to as *examples*, as they constitute the input data to be learned from, and individual variables (coordinates of the examples) are referred to as *features*.

In classification, the label set $\mathcal{Y}$ is a finite set of categories. The most common example is that of *binary* classification, where $\mathcal{Y} = \{-1, 1\}$. More recently, there has been a large body of work on *multi class* classification (Bengio et al., 2010), where $\mathcal{Y} = \{1, \dots, K\}$ is a possibly large set of categories, and on *structured prediction* (Bakir et al., 2007), where $\mathcal{Y}$ is a finite set usually induced by some combinatorial structure (such as a set of spanning trees or perfect matchings of a given tree). At the intersection between multi class and structure, *multi label* classification consists in $\mathcal{Y} = \mathcal{P}(\{1, \dots, K\})$, where $\mathcal{P}$ is the power set. In this last setting, each example may be associated with any number of the K labels. In regression, $\mathcal{Y}$ is a continuous space, such as the real line. Although classification is concerned with predicting binary values, it is usually more practical and more interesting to associate this prediction with a real-valued confidence score. We will focus on regression and on binary classification, and we will consider in both cases *predictors* of the form $f : \mathbb{R}^d \to \mathbb{R}$. For classification, the predictor provides a label through its sign, and a confidence score through its magnitude. For regression, the predictor directly estimates the real-valued target. Although providing confidence intervals for regression is an interesting problem, we do not consider it here.

2.1.1 Linear Models: Generalization and Estimation

A common choice is to use *linear* predictors of the form $f(x) = \langle x, w \rangle + b$, for $w, x \in \mathbb{R}^d$, and a bias term $b \in \mathbb{R}$. For binary classification, this consists in assuming that labels are generated according to the model

$$y_i = \varepsilon_i \operatorname{sign}(\langle x_i, w^\star \rangle + b), \text{ with } w^\star \in \mathbb{R}^d, b \in \mathbb{R}, \varepsilon \in \{-1, 1\}^N, \tag{2.1}$$

for $i \in [N]$, where ε is the *label noise*, which may flip some labels randomly. For regression, this corresponds to assuming targets following the model

$$y_i = \langle x_i, w^\star \rangle + b + \varepsilon_i, \text{ with } w^\star \in \mathbb{R}^d, b \in \mathbb{R}, \varepsilon \in \mathbb{R}^N, \tag{2.2}$$

for $i \in [N]$, where ε is the noise. We refer to the case where $\varepsilon = \mathbf{1}$ for classification (resp. where $\varepsilon = 0$ for regression) as the *noiseless* setting. Here, $\mathbf{1}$ denotes the all-ones vector. In the following and when more convenient for presentation, we will omit the bias term b, and state models for regression, although similar models for binary classification can be obtained. In the statistics literature, the set $\{x_1, \dots x_N\}$ of feature vectors we are testing against is referred to as the *design*. We will often use the matrix notation

$$X = \begin{bmatrix} x_1^T \\ \vdots \\ x_N^T \end{bmatrix} \in \mathbb{R}^{N \times d}.$$

The following are classical designs:

- **Fixed design.** This is the case where $\{x_1, \dots, x_N\}$ is assumed fixed and known. This case is common in practice when one has limited control over the acquisition of the data.

- **Fixed orthogonal design.** A special case of fixed design is the case where $X^T X = I_d$, where I_d is the identity of $\mathbb{R}^d$. When the elements $\{x_1, \dots, x_N\}$ of the design are unit-Euclidean length, they are also referred to as a *tight frame* of $\mathbb{R}^d$. In particular, for $N = d$ and $X = I_d$, observations are simply a noisy version of the parameters of the model: for $i \in [N]$,
$$y_i = w_i^\star + \varepsilon_i.$$
The corresponding regression problem is also referred to as *denoising*.

- **Random design.** This is the case where the design is randomly chosen according to some distribution. This random choice may or may not be in the control of the experimenter. In any case, the experimenter will almost always have budget restrictions on the number of observations that can be acquired.

The linear framework is actually quite general and powerful, as higher dimensional vector representations $(\bar{x}_i)$ can be devised from the initial vector representations (x_i) of the examples. This can be used to introduce nonlinearities.

A central problem in statistics and machine learning is to devise and analyze procedures which can learn with the smallest amount of examples, possibly in the presence of noise. This minimal number of examples needed to learn within a particular model is called the *sample complexity*. What should *learn* mean here? We distinguish two types of objectives:

- **Generalization.** In many cases, the objective is to be able to predict targets (labels or continuous values) on new unseen data. In this context, one does not necessarily seeks to estimate $w^\star$ directly, but pursues the potentially easier objective of making accurate predictions. For this problem, a method needs to produce a *predictor* of the form $f : \mathbb{R}^d \to \mathbb{R}$.

- **Estimation.** A harder task consists in estimating the original parameters $w^\star$ of the model. This estimate can in turn be used to make predictions, but may also be interesting in its own right to gain insight on the data. For arbitrary designs, this may not be possible, while for orthogonal designs, this is equivalent to the generalization problem. For this task, a method is required to produce an *estimator* $\widehat{w} \in \mathbb{R}^d$.

These two problems are of course closely related, and although we will mostly use the *estimation* terminology, both objectives should be kept in mind. The following problem is a classical example of estimation problem.

Exemple 1 (Compressed sensing)**.** *In compressed sensing, the model is of the form*

$$y = Xw^\star, \quad \{i : w_i^\star \neq 0\} = k.$$

This can be extended to a noisy model. The distinctive feature of this model is that it is sparse: only k coefficients of the model are nonzero.

As was already hinted at, how well one can succeed in these two types of problems is highly dependent on the true model dimension (if any) and on the amount of observations available. A general method to handle such problems is *empirical risk minimization* (ERM), wherein one defines a risk criterion directly from the observations. The traditional way to approach regression problems is through through least-squares, where one selects

$$\widehat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{i=1}^{N} (y_i - \langle x_i, w \rangle)^2.$$

This coincides with the maximum likelihood estimator of w given (y_i) when the noise vector ε is a standard Gaussian vector. We distinguish different regimes:

- When $N \geq d$, there may exist a unique solution with loss zero, which can then be obtained in closed form. When this is not the case (the linear system is overdetermined), there still exists a solution to the least-squares problem that can be obtained in closed form, although the loss is not zero.
- When $N < d$ (the associated linear system is underdetermined), there is an infinity of solution. Among these solutions, the *pseudo-inverse* allows to construct the one with the minimum Euclidean norm.

For classification, the situation is already nontrivial even if $N \geq d$. The most natural estimator consists in minimizing the *classification loss* (or, *zero-one loss*):

$$\widehat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \sum_{i=1}^{N} \mathbf{1}[y_i \neq \operatorname{sign}\langle x_i, w \rangle],$$

where $\mathbf{1}[C]$ is the indicator function that is one when condition C is true, and zero otherwise. Although this is arguably a good estimator at least in the noiseless case when $N \geq d$, the computation of the estimator reduces to a highly non-convex optimization problem in w that cannot usually be solved. In addition, such hard zero-one costs for misclassification can deteriorate the predictor in the presence of label noise. For these reasons, one almost always work with different real-input and real-output loss functions $\ell : \mathcal{Y} \times \mathbb{R}^d \to \mathbb{R}_+$ instead of the classification loss: here, $\ell(y, t)$ measures the cost of predicting a target (or, in classification, a confidence level) $t \in \mathbb{R}$ while the true target is y. Such loss functions are often referred to as *surrogates*, as they are to be minimized in place of some original loss.

In the regime where $N \geq d$, the generalization and estimation problems are well understood for both regression and classification. Difficulties there mostly arise when considering how to efficiently obtain the estimators, possibly in the presence of a lot of data. The challenges in this case are hence mostly about optimization and systems engineering.

The regime where $N < d$ is usually referred to as the *high-dimensional* setting. This regime differs from classical statistics in that when considering asymptotics, the dimension of the model is usually assumed to diverge at the same time that the number of observations (i.e., $N \to \infty$ and $d \to \infty$), such that the problem does not automatically get easier with more observations. In this case, even what is a good predictor is not necessarily clear. In general, recovering a high-dimensional model with very few observations is of course doomed in advance, and one has to come up with ways to select predictors amongst the many which may reasonably fit to the data.

2.1.2 Occam's Razor and Minimum Description Length

A general principle is *Occam's razor*, wherein the *simplest* predictor or estimator which allows to approximate the data should be preferred. According to this principle, the simplest model is deemed the most able to generalize to new unseen examples, while very complex models are deemed to have over fitted to the training data. This informal principle leaves unspecified how to measure simplicity (or complexity) of a predictor, and how to balance this with the fit to the data. This has been formalized in the *minimum description length* principle (Grünwald, 2007), where models which can be *best compressed* should be preferred. This principle is at the root of most penalized methods, which vary in how compressibility is measured.

A very early measure of compressibility is the Kolmogorov complexity (Li and Vitányi, 2009). Consider a programming language in which any universal Turing machine can be implemented, such as the C programming language, and the binary sequence

$$S = [01010101 \dots 01010101] \in \{0,1\}^{1000}.$$

The Kolmogorov complexity of S is the length of the shortest program that prints the sequence and halts. A trivial program just prints the full sequence:

```
printf("01010101...01010101");
```

However, this can be compressed much more through:

```
for (i = 0; i < 500; ++i) printf("01");
```

This allows to define complexity for predictors as well. In spite of its important historical role, Kolmogorov complexity can seldom ever be computed in practice, and has to be replaced by other measures of compressibility.

Another information theoretical notion of complexity is Shannon's mutual information (Cover and Thomas, 2012). Unlike the Kolmogorov complexity, this is a measure of complexity over random objects. Consider a discrete input random variable $X \in \mathcal{X}$, and a discrete output random variable $Y \in \mathcal{Y}$, with joint distribution $P_{X,Y}$ and marginal distributions P_X and P_Y. The mutual information between X and Y is

$$I(X;Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} P_{X,Y}(x,y) \log\left(\frac{P_{X,Y}(x,y)}{P_X(x)P_Y(y)}\right).$$

In particular, $I(X;Y)$ is zero when X and Y are independent, and is maximal when $X = Y$, where it is equal to the Shannon entropy of X. The mutual information has many interesting properties, and is in particular invariant with respect to relabelings of $\mathcal{X}$ and $\mathcal{Y}$. Informally, $I(X;Y)$ measures how predictable Y is, given that X is known, or vice-versa. Equipped with mutual information, one may seek to construct a proxy random variable T to be used for predicting Y as follows: pick T such that it compresses the information in X, while retaining the maximum amount of information about Y. The complexity of the proxy predictor T is measured here using $I(X;T)$: when the mutual information between X and T is low, the predictor T only encode a small subset of the information of X, and can thus be deemed to be compression of X. This measure of complexity is notably used

in the Information Bottleneck (IB) method (Tishby et al., 1999), where, for some parameter β, one seeks to minimize

$$\mathcal{C}_{\mathrm{IB}}(T) = I(X;T) - \beta I(Y;T),$$

with respect to the distribution of the random variable T. This objective is the Lagragian associated to a problem of the form

$$\min_{T} -I(Y;T) \text{ s.t. } I(X;T) \leq \gamma,$$

where ones aims to compress information in the example X under a constraint on how well the target Y can be predicted. In penalized estimation terminology, $-I(Y;T)$ is a *loss function*, while $I(X;T)$ is a *penalty*. This type of tradeoff will be central to all penalized estimation methods that we will cover. In principle, the optimization is over all distributions $p_{T|X}$ and p_T. Although this provides an interesting unifying framework for learning and coding, this is usually not practical, unless in special cases such as with the *Gaussian Information Bottleneck* (Chechik et al., 2005).

2.1.3 Priors and Penalized Estimation

We focus in this thesis on measuring complexity of predictors or estimators through convex functions. Formally, this corresponds to selecting a predictor using a rule of the form

$$\widehat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \left[\sum_{i=1}^{N} \ell(y_i, f(x_i)) + \lambda \Omega(f) \right],$$

where $\ell : \mathcal{Y} \times \mathcal{F} \to \mathbb{R}_+$ is the *loss function*, λ is the regularization parameter, and $\Omega : \mathcal{F} \to \mathbb{R}_+ \cup \{\infty\}$ is the *penalty* (or, *regularizer*). Linear predictors correspond to $\mathcal{F} \subset \{x \mapsto \langle x, w \rangle : w \in \mathbb{R}^d\}$, and one can directly define the regularizer as a function of the hyperplane through $\Omega(f) = \Omega(w)$ for some $\Omega : \mathbb{R}^d \to \mathbb{R}_+ \cup \{\infty\}$. In the following, we will almost always consider regularizers which are norms, and we will denote the corresponding unit ball (resp. unit sphere) by $\mathbf{B}_\Omega$ (resp. by $\mathbf{S}_\Omega$). When Ω is a ℓ_p-norm for some p, we will write for simplicity $\mathbf{B}_p$ and $\mathbf{S}_p$, such that $\mathbf{B}_2$ is the unit-ball for the Euclidean norm. A simple example of penalized estimation is ℓ_2-penalized classification or regression, which considers linear predictors of the form

$$f(x) = \langle \widehat{x}, w \rangle, \text{ with } \widehat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \left[C \sum_{i=1}^{N} \ell(y_i, \langle x_i, w \rangle) + \frac{\|w\|_2^2}{2} \right]. \tag{2.3}$$

As already mentioned, the pseudo-inverse allows to construct the least squares solution to a linear system with the minimum Euclidean norm. This can be interpreted as a sort of ℓ_2-regularization of the least squares loss, although with $C = \infty$, such that there is no way to balance the fit to the data and the ℓ_2-norm complexity. The following is a classical example of method based on a ℓ_2-norm penalty for regression.

Exemple 2 (Regression: Ridge regression)**.** *Ridge regression corresponds to the case of the squared loss in Equation* (2.3)*:*

$$\ell_2(y,t) = (y-t)^2.$$

This is equivalent to standard least squares on a regularized Gram matrix (Tibshirani et al., 2009), and is usually more used for regression, although this can in principle be used for classification.

The following are examples for classification, where the primary objective is usually to obtain good generalization performance.

Exemple 3 (Classification: SVMs)**.** *The case where the loss in Equation (2.3) is the so-called hinge loss*

$$\ell_{hinge}(y,t) = (1-yt)_+$$

is referred to as a Support Vector Machines (SVMs) (Boser et al., 1992; Steinwart and Christmann, 2008).

Exemple 4 (Classification: Penalized logistic regression)**.** *Penalized logistic regression is another method for classification with the ℓ_2-norm penalty, which uses the logistic loss in Equation (2.3):*

$$\ell_{logit}(y,t) = \log(1+\exp(-yt)).$$

With ℓ_2-norm penalization, the resulting objective function is twice differentiable everywhere, which allows to leverage second-order optimization methods.

A common way to view penalized estimation is that of *priors* as in Bayesian statistics. Assume that you have some prior information on the object to estimate (e.g., on its support, sign, sparsity, maximum magnitudes, etc.). In this case, one may use penalized estimation to try to find a predictor with the desired properties, by designing a regularizer Ω which favors such predictors. In many situations, a probabilistic model can be defined from a penalized objective. Informally, this consists in making additional hypotheses on the model, thus making it easier to estimate. For instance, even if $N < d$, a linear model may actually be straightforward to estimate if you know in advance that only a single or a small number of features are useful. Two main types of such additional hypotheses on models have both had an immense influence in the past years in machine learning: sparsity, and latent factors. Section 2.2 and Section 2.3 are devoted to exploring in depth these two hypotheses, and associated regularizers. However, before going in detail into these ideas, we generalize slightly the linear models that we consider.

2.1.4 Penalized Matrix Estimation

In this part of the thesis, we are mostly interested in the estimation of matrices, and we actually consider more general matrix linear models of the form

$$y_i = \langle X_i, W^\star \rangle + \varepsilon_i, \text{ with } W^\star \in \mathbb{R}^{d_1 \times d_2}, \tag{2.4}$$

for design matrices $X_i \in \mathbb{R}^{d_1 \times d_2}$, and $i \in [N]$. As previously, we refer to the collection $\{X_1, \dots, X_N\}$ as the *design*. Most priors on vectors can be extended in straightforward ways to priors on matrices. For instance, consider the problem of learning to classify with K classes (i.e., a multi class problem). This can be reduced to K binary classifiers (each of them classifying a given class against the rest of the classes), in the form of a matrix $W \in \mathbb{R}^{K \times d}$. However, viewing coefficients as a rectangular array opens up new possibilities to take into account more structure.

Exemple 5 (Covariance matrix estimation). *Consider the task of estimating the covariance matrix of a d-dimensional vector (Karoui, 2008; Cai et al., 2010), from observations $x_1, \dots, x_N \in \mathbb{R}^d$. The usual estimator is the sample covariance matrix*

$$\widehat{\Sigma} = \frac{1}{N} \sum_{i=1}^{N} x_i x_i^T \in \mathbb{R}^{d \times d}.$$

When $N < d$, this estimator is singular, and further hypotheses may allow to obtain a better estimator. For instance, there may only be a small number of variables (features) interacting with each other, and penalized estimation may allow to define an estimator that takes this into account. The estimator $\widehat{\Sigma}$ needs to be a valid covariance matrix, and hence must be a positive semi-definite matrix.

Exemple 6 (Dynamic link prediction). *Consider a graph G with a fixed vertex set V, but with a dynamically changing edge set E_t, where t denotes the time. The adjacency matrix A_t of the graph at time t is observed at subsequent time instants $t_1, \dots, t_N$, leading to a series of snapshots $A_{t_1}, \dots, A_{t_N}$ of the graph. In the simplest model, edges can only be added to the graph, although one may in principle consider a fully general evolution where edges can be both added or removed. In link prediction (Taskar et al., 2003; Liben-Nowell and Kleinberg, 2007), the objective is to predict from the snapshots what the next state of the graph will be: produce an estimator $\widehat{A_{t_{N+1}}}$ of the next adjacency matrix $A_{t_{N+1}}$. This can be formulated as a regression/classification problem over a symmetric binary (or weighted) matrix.*

This last example can actually be expressed in the more general framework of *matrix completion*.

Exemple 7 (Matrix completion). *In matrix completion (Candès and Recht, 2009), the observation is an incomplete version of a matrix $W^\star \in \mathbb{R}^{d_1 \times d_2}$. The model is of the form*

$$Y = \Omega \circ W^\star$$

where $\Omega \in \{0,1\}^{d_1 \times d_2}$ is a known observation mask, and the Hadamard product is defined for two matrices X and Y as $(X \circ Y)_{i,j} = X_{i,j} Y_{i,j}$. This can be extended to include noise in the observations, and more complicated observation masks or designs (Koltchinskii et al., 2011b).

These three problems will be discussed further in Chapter 3. We will consider a matrix estimation problem related to *multiple kernel learning* in Chapter 4. Many other problems, such as multi-task learning (Evgeniou and Pontil, 2004), can be cast as matrix problems. Similarly as in the vector case, matrix predictors can be obtained through penalized problems of the form

$$f(X) = \langle X, \widehat{W} \rangle, \text{ with } \widehat{W} = \operatorname*{arg\,min}_{W \in \mathbb{R}^{d_1 \times d_2}} \left[\sum_{i=1}^{N} \ell(y_i, \langle X_i, W \rangle) + \lambda \Omega(W) \right],$$

where $\Omega : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_+ \cup \{\infty\}$ is a matrix regularizer.

2.1.5 Penalized or Constrained?

A variation on penalized estimation is *constrained* estimation, where a hard-constraint is used instead of the penalty:

$$\widehat{f} = \operatorname*{arg\,min}_{f \in \mathcal{F}} \sum_{i=1}^{N} \ell(y_i, f(x_i)) \text{ s.t. } \Omega(f) \leq \gamma.$$

Unless γ is large enough, the predictor will satisfy $\Omega(\widehat{f}) = \gamma$ (that is, the constraint is active). This allows to precisely constraint the complexity of the predictor as measured by Ω. However, from Lagrangian duality, any such constrained program can be reformulated as a penalized estimation problem with a regularization parameter $\lambda = \lambda(\gamma, (x_1, y_1), \ldots, (x_N, y_n))$. This mapping is usually unknown and data-dependent, and, in both cases, the hyper-parameter λ or γ usually has to be tuned empirically. Depending on the situation, one or the other formulation may be more amenable to optimization or theoretical analysis. The family $(\widehat{f})_{\gamma \in \mathbb{R}_+}$ of all estimators which can be obtained by varying the hyper-parameter is referred to as the *regularization path*. In some cases, full regularization paths can be obtained in a finite number of steps for penalized estimators. The following characterization is from Rosset and Zhu (2007).

Proposition 2. *Consider the estimator* $\widehat{w} = \operatorname{arg\,min}_{w \in \mathbb{R}^d} [L(w) + \lambda \Omega(w)]$ *with* $\lambda \geq 0$ *and* L, Ω *convex over* $\mathbb{R}^d$. *If* L *is piecewise quadratic and* Ω *piecewise linear, the estimator has a piecewise linear regularization path.*

This applies, in particular, to SVMs (Hastie et al., 2004) or Lasso (Efron et al., 2004), for which the full regularization path can be obtained at a cost of the same order as that of obtaining the estimator for a single parameter (provided that the piecewise linear path has only a constant number of segments). This does not apply to many other models, such as Ridge regression which has polynomial paths as illustrated in Figure 2.1, although other techniques can be used to efficiently approximate smooth paths over a finite grid of hyper-parameters (Bach et al., 2005). However, for many methods, there is no low-cost way of obtaining regularization paths, and how sensitive a method is to hyper-parameters is a paramount element when comparing penalized methods.

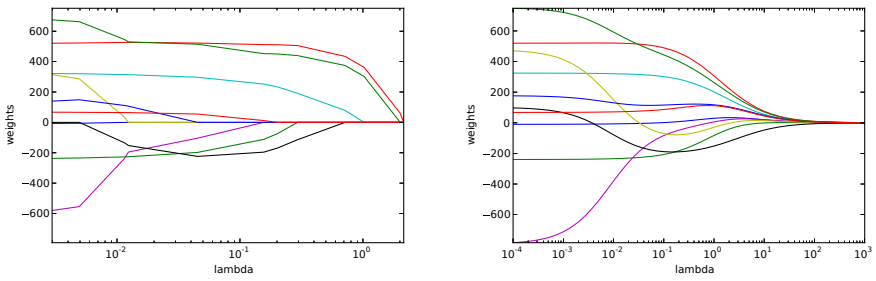

Figure 2.1: Example of regularization paths for Lasso (left) and Ridge regression (right): each curve corresponds to a coordinate of the solution $w^\star$

2.2 Sparsity

The notion of sparsity has played a paramount role in machine learning and statistics over the last decades, and we study the associated concepts and ideas in this section. For a linear predictor $f(x) = \langle x, w \rangle$, sparsity of w can be interpreted as a *variable selection*, where only a small subset of all available variables are deemed useful for prediction. Performing feature selection has many advantage: corresponding predictors are usually associated with lower generalization errors than full models, are amenable to theoretical analysis, and can be more computationally efficient both at training and testing. The *sparsity* terminology is often used with different meanings. A matrix (or vector) is considered sparse if it has few nonzero elements. On the other hand, it is common to refer to the number of nonzero elements of a matrix as its sparsity, in a slight abuse of language. In the context of sparse methods, the *sparsity pattern* refers to the support.

2.2.1 The ℓ_0-norm

The problem of selecting variables (in a binary fashion) can be cast as constraints or regularizers based upon the size of the support of the vector to estimate, which we refer to as the ℓ_0-norm:

$$\|w\|_0 = \operatorname{card}\{i : w_i \neq 0\}.$$

Although we use the *norm* terminology, $\|.\|_0$ is not actually a norm, as it does not separate points and it is not positively homogeneous. The use of the ℓ_0-norm traces back to greedy methods (Mallows, 1973; Akaike, 1974; Mallat and Zhang, 1993; Efron et al., 2004) which at every step add (or, possibly, remove) a single coordinate to or from the model , until reaching the desired sparsity. This includes forward stage wise selection and related methods which can produce full regularization paths.

The Iterative Hard Thresholding (IHT) method is geared towards optimization problems of the form

$$\min_{w \in \mathbb{R}^d} L(w) \text{ s.t. } \|w\|_0 \leq k$$

for a loss $L : \mathbb{R}^d \to \mathbb{R}$. IHT corresponds to projected gradient descent with respect to the loss L, and the non-convex constraint set $\mathbf{B}_0 = \{w : \|w\|_0 \leq k\}$. In particular, for sparse least-squares regression with $L(w) = \|Xw - y\|_2^2$ and $\|X\|_{\text{op}} \leq 1$, this is theoretically guaranteed to converge to a local minimum (Blumensath and Davies, 2009), and IHT consists in this case in the iteration

$$w_{n+1} = \text{HT}_k(w_n + X^T(y - Xw_n)),$$

where HT_k is the hard thresholding operator that keeps only the largest k elements (in magnitude) of a vector, and set the remaining coordinates to zero. Unlike greedy methods, IHT does not build up a sparse vector by iteratively adding new nonzero coordinates, but maintains a k-sparse vector throughout.

In general, however, optimization problems involving the size of the support often are combinatorial and NP-hard to solve to a global optimum. In addition, it may not always be desirable to measure complexity through the size of the support. In practice and when computing with floating-point numbers, a threshold needs to be set to control when a coefficient is deemed small enough to be zero. In all cases, this measure of complexity is highly unstable under arbitrarily small energy perturbations. These reasons have prompted the

development of other means for measuring to what extent a vector is concentrated over a few coordinates only.

2.2.2 The ℓ_1-norm

A popular approximation (Chen and Donoho, 1994; Tibshirani, 1996b; Chen et al., 1998) consists in substituting the size of the support with the ℓ_1-norm, defined for $w \in \mathbb{R}^d$ as

$$\|w\|_1 = \sum_{i=1}^{d} |w_i|.$$

In addition to leading to usually tractable convex optimization problems, the ℓ_1-norm as a measure of complexity alleviates the unstability of the size of the support. The ℓ_1-norm penalty is not differentiable at any point where $\|w\|_0 \neq d$, which prevents the use of second-order methods out of the box. However, it does admit a sub differential everywhere, which is for any $i \in [d]$

$$[\partial \|.\|_1]_i (w) = \begin{cases} \operatorname{sign}(w_i) & \text{if } w_i \neq 0, \\ [-1, 1] & \text{if } w_i = 0. \end{cases}$$

This allows to perform optimization with the ℓ_1-norm, as we will show in Chapter 3. The ℓ_1-norm admits various variational expressions. Duality between norms provide the following:

$$\|w\|_1 = \sup \{\langle z, w \rangle : \|z\|_\infty \leq 1\},$$

where the supremum is attained at $z = (\operatorname{sign} w_1, \ldots, \operatorname{sign} w_d)$. Another variational expression can be derived from the Cauchy-Schwarz inequality, leading to

$$\|w\|_1 = \frac{1}{2} \min \left\{ \sum_{i=1}^{d} \left[\frac{w_i^2}{z_i} + z_i \right] : z \in \mathbb{R}_+^d \right\}.$$

A simple everywhere differentiable approximation to the ℓ_1-norm is

$$\|w\|_{1,\varepsilon} = \sum_{i=1}^{d} \sqrt{|w_i|^2 + \varepsilon^2}.$$

This penalty retains the linear rate of increase away from zero, and also allows to avoid non-differentiability issues. In practice, this later penalty is rarely used, probably due to the availability of efficient non-differentiable optimization methods (such as proximal methods), due to good performance of everywhere differentiable regularizers such as the ℓ_2-norm, or because this introduces the extra parameter ε.

Exemple 8 (Lasso). *The Lasso (Tibshirani, 1996b) is the estimator*

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^d} \left[\|Xw - y\|_2^2 + \lambda \|w\|_1 \right].$$

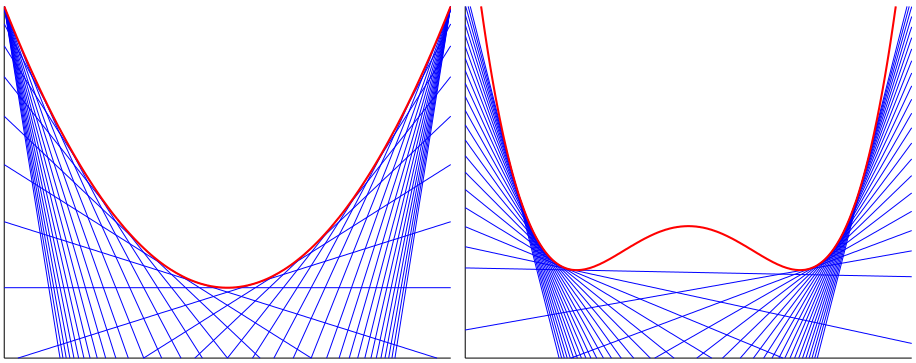

Figure 2.2: Illustration of convex conjugacy on a convex quadratic (left), non-convex function (right): original function (red), supporting hyperplanes to epigraph (blue)

The choice of the ℓ_1-norm for measuring sparsity can be derived from a geometrical perspective. In words, the ℓ_1-norm is the largest convex lower bound to the ℓ_0-norm over the ℓ_∞-norm unit ball. Indeed, notice that $\|w\|_1 \leq \|w\|_\infty \|w\|_0$ for $w \in \mathbb{R}^d$. The ℓ_1-norm can be shown to be the largest such lower bound. This can be formalized through convex duality.

Definition 1. *Consider $f : \mathbb{R}^d \to \mathbb{R} \cup \{\infty\}$. The convex conjugate of f is defined for $z \in \mathbb{R}^d$ as*

$$f^\star(z) = \sup_{x \in \mathbb{R}^d} \left[\langle x, z \rangle - f(x) \right].$$

The conjugate of a function is a description of the function not in terms of function values in the original space, but in terms of all the supporting hyperplanes to the epigraph. For convex lower semi-continuous functions, these two descriptions are equivalent (Rockafellar, 1997).

Proposition 3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and lower semi-continuous, and denote by $f^{\star\star}$ the biconjugate of f defined as $(f^\star)^\star$. Then, $f = f^{\star\star}$.*

This property is not true for non-convex functions, as illustrated in Figure 2.2. However, for non-convex functions, it holds that $f^{\star\star} \leq f$, and the convex biconjugate is always a lower bound of f. In fact, by Proposition 3, the biconjugate is necessarily the *largest convex lower bound* to f. This can be used to find a *convex surrogate* or *convex envelope* to the ℓ_0-norm. To simplify notation, define the barrier indicator function of a set $X \subset \mathbb{R}^d$ by

$$\delta_X(w) = \begin{cases} 0 & \text{if } w \in X, \\ \infty & \text{otherwise.} \end{cases}$$

Consider the following restriction of the ℓ_0-norm: $f(w) = \|w\|_0 \, \delta_{\|.\|_\infty \leq 1}(w)$. Then,

$$f^\star(z) = \sum_{i=1}^d \sup_{|w_i| \leq 1} \left[w_i z_i - \mathbf{1}(w_i \neq 0) \right] = \sum_{i=1}^d (z_i - 1)_+.$$

The biconjugate is thus

$$f^{\star\star}(w) = \sum_{i=1}^{d} \sup_{z_i} [w_i z_i - (z_i - 1)_+] = \begin{cases} \|w\|_1 & \text{if } \|w\|_\infty \leq 1, \\ \infty & \text{otherwise.} \end{cases}$$

Hence, over bounded vectors, the ℓ_1-norm is the *tightest* convex relaxation to the ℓ_0-norm. Note that it is necessary here to fix the scale by looking at the unit ball for the ℓ_∞-norm: the convex envelope to the ℓ_0-norm over $\mathbb{R}^d$ is the uniformly zero function.

2.2.3 Elastic-net and Variations

This rationale for using the ℓ_1-norm has been criticized in various ways, and one of them is related to how it behaves in terms of correlated variables. Consider the case where some relevant variables are very much correlated. The ℓ_1-norm penalty may only select a single of these variables, and which variable is selected might be numerically unstable. In particular, consider the case where features 1 and 2 are always equal: for any given C, any ℓ_1-norm penalized objective in dimension two has the same value over all of $\left\{w \in \mathbb{R}^2 : w_1 + w_2 = C\right\}$. The elastic net penalty (Zou and Hastie, 2005b) was introduced in part to remedy this problem, and consists in

$$\|w\|_{\text{elastic}} = \|w\|_1 + \gamma \|w\|_2^2.$$

The strict convexity of the ℓ_2-norm part encourages joint selection of correlated features. In the previous examples of equal features,

$$\|(C,0)\|_{\text{elastic}} = C + \gamma C^2 > \|(C/2, C/2)\|_{\text{elastic}} = C + \frac{\gamma}{2} C^2,$$

such that the elastic net encourages configurations where both features are equally used. The ℓ_2-norm penalty shrinks the estimator, and the resulting w is usually rescaled by $1+\gamma$ to make up for this. The Trace Lasso penalty (Grave et al., 2011) induces a similar effect in a data dependent matter, through a spectral technique:

$$\|w\|_{\text{TL}} = \|X \operatorname{diag}(w)\|_*,$$

where $\|.\|_*$ is the trace norm, which is the sum of the singular values. Informally, this measures the dimension of the linear subspace spanned by the selected features. In particular, the penalty does not increase when selecting additional features very much correlated with a feature that was already selected. Unlike most regularizers, the Trace Lasso penalty measures complexity of the predictor w based on the knowledge of the data. This is similar in principle to the mutual information regularizer $I(X,T)$ from the Information Bottleneck.

The k-support norm $\|\cdot\|_k^{\text{sp}}$ (Argyriou et al., 2012) corresponds to the tightest convex surrogate of the ℓ_0-norm, but over the Euclidean unit ball instead of the ℓ_∞-norm unit ball. This can be formulated in a slightly different framework than before. Instead of taking the point of view of *convex functions*, where one looks for the largest convex lower bound, consider that of *convex sets*, where one looks on the opposite for the smallest convex superset of a non-convex subset (also referred to as the *convex hull*), as illustrated in Figure 2.3. For the k-support norm, this can be stated as

$$\operatorname{conv}\{w : \|w\|_0 \leq k, \|w\|_2 \leq 1\} = \left\{w : \|w\|_k^{\text{sp}} \leq 1\right\},$$

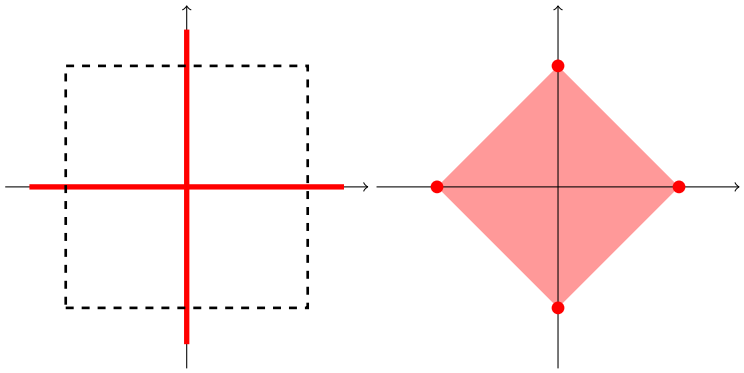

Figure 2.3: Convex set view of ℓ_1-norm relaxation: ℓ_∞-norm unit ball (dashed black), 1-sparse vectors (red), intersection of both (red points), convex hull of intersection (light red)

where conv is the convex hull operator . There is a noteworthy difference with the ℓ_1-norm relaxation, which in similar terms can be written

$$\operatorname{conv}\{w : \|w\|_0 \le k, \|w\|_\infty \le 1\} = \left\{w : \frac{1}{k}\|w\|_1 \le 1\right\}.$$

In the ℓ_1-norm case, the convex relaxation depends in an homogeneous fashion in the number k of nonzero elements, while with the k-support norm, this dependence is more intricate. Although the relaxation result for the ℓ_1-norm can be equally easily formulated in terms of convex biconjugate or convex hulls, the k-support norm comes up most naturally with convex hull. The k-support norm is closely related to the following constraint reformulation of the elastic net

$$\|w\|_k^{\text{elastic}} = \max\left\{\|w\|_2, \frac{\|w\|_1}{\sqrt{k}}\right\},$$

which is such that $\|w\|_k^{\text{elastic}} \le \|w\|_k^{\text{sp}} \le \sqrt{2}\|w\|_k^{\text{elastic}}$ for $w \in \mathbb{R}^d$. In spite of this, the k-support norm has been shown to outperform the elastic net in a variety of settings.

2.2.4 Other Regularizers

Many other notions of sparsity have been proposed, and we only mention a few. Non-uniform weightings over the features can be used, and some interesting such weightings are *adaptive*: they are chosen directly from the data. A simple example is the adaptive Lasso (Zou, 2006): given an initial estimate $\widehat{w}$, consider the penalty

$$\|w\|_{1,\widehat{w}} = \sum_{i=1}^{N} \frac{|w_i|}{|\widehat{w}_i|^\gamma}$$

for some γ. This procedure was shown to have theoretical guarantees superior to that of Lasso in terms of estimation and recovery of the support of $w^\star$. This type of reweighting can also be iterated multiple times: the estimator obtained using an adaptive Lasso-like penalty can be used to define a weighting for a new penalty, and so on.

The previous regularizers are mostly convex, but non-convex regularizers have been studied as well. The sparseness measure (Hoyer, 2004b) consists in an affine transformation of

$$\|w\|_{1,\text{eff}} = \frac{\|w\|_1}{\|w\|_2},$$

which is referred to as the *effective sparsity* (Plan and Vershynin, 2013a). Others regularizers have been defined to limit the increase of the ℓ_1-norm away from zero. The simplest example is the capped ℓ_1-norm:

$$\|w\|_{1,\text{capped}} = \min\left[\|w\|_1, \frac{1}{2}\right].$$

This introduces more singularities though, and many variants have been proposed, such as SCAD (Fan and Li, 2001) and MCP (Zhang, 2010). Finally, note that all of these regularizers can be extended in a straightforward manner to matrices by viewing them as vectors in $\mathbb{R}^{d^2}$.

2.3 Rank and Latent Factor Models

Latent factor models are a wide class of models, based on a very different type of hypothesis than sparse models. They apply to problems with relational or high-order structures, such as in the following example.

Exemple 9 (Recommender systems)**.** *In recommender systems (Koren, 2008), one seeks to recommend products to customers. These systems are ubiquitous in e-commerce and social websites, but extend well beyond these. In the simplest setting, one has data corresponding to past purchases. When the only information available is whether a purchase was made for each user/product combination, this is link prediction as in Example 6. However, in many cases, ratings are collected to quantify how happy each user was with each product purchased. This allows to cast the problem as a matrix completion problem as in Example 7: consider the ratings matrix $W^\star$ of size (number of users) $\times$ (number of products), where for a user u and a product p,*

$$W^\star_{u,p} = \begin{cases} \textit{rating given by } u \textit{ to } p & \textit{if } u \textit{ purchased } p, \\ \square & \textit{otherwise.} \end{cases}$$

Here, the symbol $\square$ indicates that the information is unavailable and must be inferred. The objective is, for each user u, to provide a shortlist of products. In general, the problem is very much ill-posed from a mathematical standpoint (due to the absence of clear statistical model in practice), and evaluation metrics are thus an important part of a recommender system.

In Example 9, there are multiple types of entities, which interact according to some weighting W to be estimated. In latent factor models, the hypothesis is that there are a small number of *latent factors*, which regulate how entities interact. The latent factors are new kinds of entities (u_i) and (v_j) in some common vector space $\mathbb{R}^r$ and such that

$$W_{i,j} \simeq \langle u_i, v_j \rangle.$$

Formally, this means that $A \in \mathbb{R}^{d_1 \times d_2}$ can decompose as a product of matrices, or *factorizes*, as

$$A \simeq UV^T$$

where $U \in \mathbb{R}^{d_1 \times r}, V \in \mathbb{R}^{d_2 \times r}$ for some r which is the number of latent factors. In Example 9, this means that every user i has a representation u_i in $\mathbb{R}^r$, and every product j has a representation v_j in $\mathbb{R}^r$, such that the rating given by user i to product j can be modeled by the norms and angle in this common vector space. This is illustrated on the left panel of Figure 2.4. Equivalently, this can be seen as in the right panel of Figure 2.4, where the the rating given by user i to product j is modeled by the sum of all the products of values across paths of length two from i to j. The assumption that there are only a small number

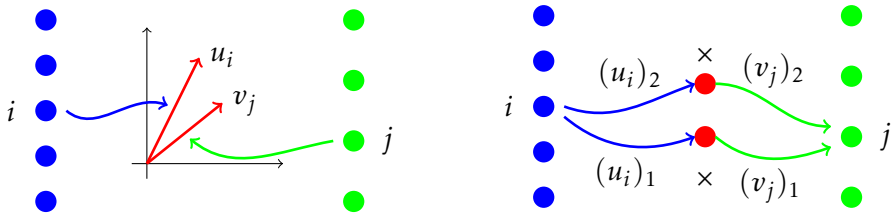

Figure 2.4: Pictorial representation of latent factor models in the context of recommender systems: users (blue), products (green), latent factors (red)

of latent factors (i.e., r is small) can greatly reduce the dimension of the problem. The usual reformulation of this hypothesis is that A should have approximately low rank. This has received a large amount of attention in the last decade, and we review in the following some of the main concepts related to it.

2.3.1 The Rank

The rank is the minimal number of elementary matrices of the form ab^T for $a \in \mathbb{R}^{d_1}, b \in \mathbb{R}^{d_2}$ which can be used to additively decompose a linear operator from $\mathbb{R}^{d_2}$ to $\mathbb{R}^{d_1}$. Formally, the rank of $W \in \mathbb{R}^{d_1 \times d_2}$ is

$$\operatorname{rank}(W) = \inf\left\{\|\sigma\|_0 : W = \sum_i \sigma_i b_i a_i^T,\ \|b_i\|_2 = \|a_i\|_2 = 1\right\},$$

with $a_i \in \mathbb{R}^{d_2}$ and $b_i \in \mathbb{R}^{d_1}$. Decompositions which achieve the infimum are given by the Singular Value Decomposition (SVD) of W, and the corresponding (a_i) (resp. the corresponding (b_i)) are mutually orthogonal. The rank is a non-convex matrix functional: in dimension two, and as illustrated in Figure 2.5, the manifold of symmetric 2×2 matrices with rank at most one is

$$\left\{\begin{bmatrix} a & b \\ b & c \end{bmatrix} : ac = b^2\right\},$$

which is simply the union of two cones. In higher dimensions, more polynomial equations are required to describe the corresponding manifolds, and the actual geometry becomes intricate very quickly.

2.3.2 The Trace Norm

Denote by $\sigma(W) = \left(\sigma_1(W), \ldots, \sigma_{\operatorname{rank}(W)}(W)\right)$ the singular values of $W \in \mathbb{R}^{d_1 \times d_2}$. The trace norm is the spectral analogue of the ℓ_1-norm.

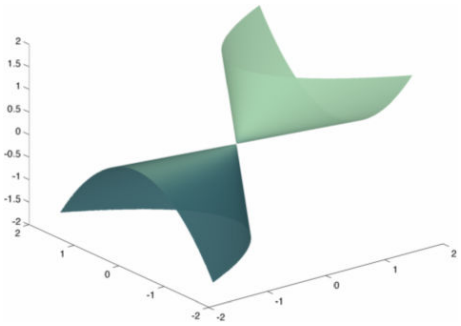

Figure 2.5: Manifold of symmetric 2×2 matrices with rank at most one

Definition 2. *The trace norm of* $W \in \mathbb{R}^{d_1 \times d_2}$ *is*

$$\|W\|_* = \sum_{i=1}^{\text{rank}(W)} \sigma_i(W).$$

In particular, $\|W\|_* = \|\sigma(W)\|_1$ and when W is positive semi-definite (which we will write $W \succeq 0$), the trace norm is simply the trace: $\|W\|_* = \text{Tr}(W)$. This norm is a special case, corresponding to $p = 1$ in the Schatten p-norms family, defined as

$$\|W\|_{*,p} = \left[\sum_{i=1}^{d} \sigma_i(W)^p\right]^{1/p}.$$

These norms are invariant with respect to unitary transformations, and are sub-multiplicative. As with classical ℓ_p-norms, the dual norm to $\|\cdot\|_{*,p}$ is $\|\cdot\|_{*,q}$, where $1/p + 1/q = 1$. In the case of the trace norm, the dual norm is the operator norm

$$\|W\|_{\text{op}} = \max_i \sigma_i(W).$$

As shown by Fazel (2002), the trace norm is the tightest convex relaxation to the rank over the operator norm ball. This is illustrated in Figure 2.6. Here, we show unit balls for the

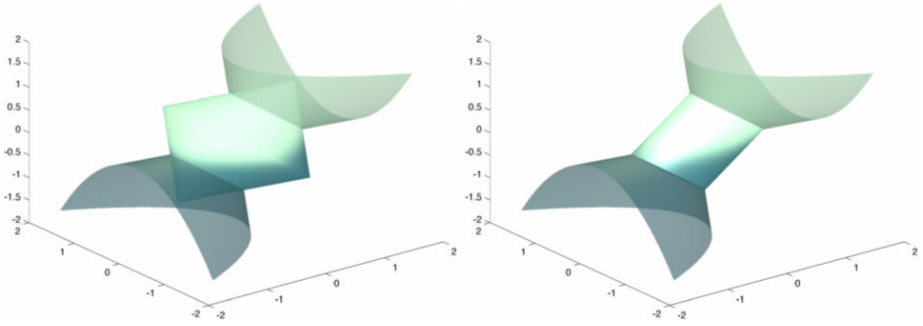

Figure 2.6: Operator norm ball (left), Trace norm ball (right)

operator norm (which is diamond-shaped) and the trace norm (which is cylinder-shaped) relaxation in the space of 2×2 symmetric matrices, with the manifold of rank 1 matrices in

overlay. In particular, the unit ball of the trace norm is the convex hull of the intersection between the rank one manifold and the operator unit ball. Formally, the fact that this is the tightest relaxation can be shown similarly as with the ℓ_1-norm relaxation of the ℓ_0-norm. Consider

$$f(W) = \mathrm{rank}(W)\ \delta_{\|\cdot\|_{\mathrm{op}}\leq 1}(W).$$

Using the equality case in the von Neumann trace theorem allows to transport the problem into the space of singular values, where the calculation is identical as for the ℓ_0-norm:

$$f^\star(Z) = \sup_{X\in\mathbb{R}^{d_1\times d_2}} [\langle X, Z\rangle - f(X)] = \sum_{i=1}^{\min(d_1,d_2)} (\sigma_i(Z)-1)_+.$$

As a consequence,

$$f^{\star\star}(W) = \|W\|_*\ \delta_{\|\cdot\|_{\mathrm{op}}\leq 1}(W).$$

As previously, one may consider the point of view of convex sets instead of convex functions. In this framework,

$$\mathrm{conv}\Big\{W : \mathrm{rank}(W)\leq r, \|W\|_{\mathrm{op}}\leq 1\Big\} = \Big\{W : \frac{1}{r}\|W\|_*\leq 1\Big\}.$$

As with ℓ_1-norm, this is homogeneous in the rank r. From a geometric point of view, the associated unit ball $\mathbf{B}_*$ is generated by unit-norm rank one matrices:

$$\mathbf{B}_* = \mathrm{conv}\Big\{uv^T : \|u\|_2 = \|v\|_2 = 1\Big\}.$$

As a consequence, $\mathbf{B}_*$ has an infinite number of extremal points. In spite of this and perhaps remarkably, projection onto $\mathbf{B}_*$ can be achieved in time $O(d^3)$ through a singular value decomposition. The trace norm can also be formulated as

$$\|W\|_* = \inf_{UV^T=W} \|U\|_F\|V\|_F = \frac{1}{2}\inf_{UV^T=W}\Big[\|U\|_F^2 + \|V\|_F^2\Big].$$

In this formulation, the dimensions of U and V may be left unconstrained, although the infimum will be attained at elements corresponding to the SVD of W. Through this formulation, it appears that the trace norm measures the average row norm of factorizations of the matrix.

Exemple 10 (Convex Matrix Completion)**.** *The matrix completion problem can be addressed when $W^\star$ is low rank using an estimator of the form*

$$\widehat{W} = \operatorname*{arg\,min}_{W\in\mathbb{R}^{d_1\times d_2}} \|W\|_*\ \textit{s.t.}\ W\circ\Omega = Y.$$

The theoretical performances of this estimator have been analyzed by Candès and Recht (2009).

In a matrix completion context, the trace norm regularizer can be very sensitive to the sampling distribution of the observed entries. A common theoretical model for matrix completion assumes that observed entries are sampled according to the uniform distribution (with or without replacement), but this is rarely the case in practice. Consider that one observes N entries of a matrix $A\in\mathbb{R}^{d_1\times d_2}$, where these entries are selected i.i.d. from

a distribution $\mathbf{P}$ over $[d_1] \times [d_2]$. A meaningful metric in this case is the generalization error under the sampling distribution, for instance in ℓ_2-norm. As shown by Srebro and Salakhutdinov (2010), even with respect to this metric, the trace norm regularizer may yield poor performance (formally, the sample complexity is unnecessarily high) when $\mathbf{P}$ deviates from uniform. In order to remedy this, they propose when the marginals of the sampling distribution are known to use a weighted variant of the trace norm. Denote by $\mathbf{P}_{\text{row}}$ (resp. by $\mathbf{P}_{\text{col}}$) the row marginal (resp. the column marginal), such that $\mathbb{P}_{\text{row}}$ can be viewed as a d_1-dimensional vector, and similarly for $\mathbf{P}_{\text{col}}$. The weighted trace-norm is

$$\|W\|_{\text{tr}(\mathbf{P}_{\text{row}},\mathbf{P}_{\text{col}})} = \left\| \text{diag}\left(\sqrt{\mathbb{P}_{\text{row}}}\right) X \, \text{diag}\left(\sqrt{\mathbb{P}_{\text{col}}}\right) \right\|_* .$$

In practice, these marginal distributions have to be estimated empirically. A generalization is the *local max norm* (Foygel et al., 2012): consider a set $\mathcal{R}$ (resp. a set $\mathcal{C}$) of marginal distributions over rows (resp. over columns), and define

$$\|W\|_{(\mathcal{R},\mathcal{C})} = \max_{\substack{\mathbf{P}_{\text{row}} \in \mathcal{R}, \\ \mathbf{P}_{\text{col}} \in \mathcal{C}}} \|W\|_{\text{tr}(\mathbf{P}_{\text{row}},\mathbf{P}_{\text{col}})}.$$

Note that $\mathcal{R}$ and $\mathcal{C}$ are then extra parameters, which need to be chosen properly.

2.3.3 Nuclear and Atomic Norms

In machine learning and statistics, the trace norm is also often referred to as *the nuclear norm*, although we will avoid this terminology, as this is a term from functional analysis which goes much further beyond the trace norm, which is merely *a* nuclear norm. In the following, we give some background on 1-nuclear norms for operators between finite dimensional spaces, as they provide inspiration for some interesting generalizations. More information can be found in the book by Jameson (1987).

Consider a matrix $W \in \mathbb{R}^{d_1 \times d_2}$. This can be viewed as an operator from $\mathbb{R}^{d_2}$ into $\mathbb{R}^{d_1}$. We will consider that the input space $\mathbb{R}^{d_2}$ is equipped with a norm $\|\cdot\|_A$. The dual norm is defined as

$$\|z\|_A^\star = \sup\{\langle w, z\rangle : \|w\|_A \leq 1\}.$$

Similarly, the output space $\mathbb{R}^{d_1}$ is equipped with a norm $\|\cdot\|_B$. The 1-nuclear norm is a norm from $(\mathbb{R}^{d_2}, \|\cdot\|_A)$ into $(\mathbb{R}^{d_1}, \|\cdot\|_B)$.

Definition 3. *The 1-nuclear norm from* $(\mathbb{R}^{d_2}, \|\cdot\|_A)$ *into* $(\mathbb{R}^{d_1}, \|\cdot\|_B)$ *of* W *is*

$$\|W\|_{nuc} = \inf\left\{ \left(\max_i \|b_i\|_B\right) \cdot \sum_i \|a_i\|_A^\star \, : \, W = \sum_i b_i a_i^T \right\},$$

with $a_i \in \mathbb{R}^{d_2}, b_i \in \mathbb{R}^{d_1}$.

The 1-nuclear norm is actually the dual matrix norm to the operator norm from $(\mathbb{R}^{d_1}, \|\cdot\|_B)$ into $(\mathbb{R}^{d_2}, \|\cdot\|_A)$, given for $Z \in \mathbb{R}^{d_2 \times d_1}$ by

$$\|Z\|_{B \to A} = \sup_{\|b\|_B \leq 1} \|Zb\|_A.$$

Formally, we have the following characterization.

Proposition 4. *For $Z \in \mathbb{R}^{d_2 \times d_1}$,*

$$\|Z\|_{nuc}^{\star} = \sup\left\{a^T Z b \,:\, \|a\|_A \leq 1, \|b\|_B \leq 1\right\} = \|Z\|_{B\to A}.$$

In practice, other definitions for the 1-nuclear norms can be used equivalently. For any decomposition $W = \sum_i b_i a_i^T$, a decomposition with equal cost is given by

$$W = \sum_i \frac{b_i}{\|b_i\|_B} \left(\|b_i\|_B a_i\right)^T.$$

As a consequence, the norm can be equivalently expressed as

$$\begin{aligned}\|W\|_{\text{nuc}} &= \inf\left\{\sum_i \|b_i\|_B \|a_i\|_A^{\star} \,:\, W = \sum_i b_i a_i^T\right\}. \\ &= \inf\left\{\sum_i \sigma_i \,:\, W = \sum_i \sigma_i b_i a_i^T, \|b_i\|_B = \|a_i\|_A^{\star} = 1, \sigma_i \geq 0\right\}.\end{aligned}$$

It can be seen from this representation that the unit ball of a nuclear norm is of the form

$$\mathbf{B}_{\text{nuc}} = \text{conv}\left\{ba^T \,:\, \|b\|_B = \|a\|_A^{\star} = 1\right\}.$$

We recover the following familiar example of 1-nuclear norm.

Exemple 11 (Trace norm). *The case where $\|.\|_A^{\star} = \|.\|_B = \|.\|_2$ (the Euclidean norm, which satisfies $\|.\|_2 = \|.\|_2^{\star}$) leads to the trace norm:*

$$\|W\|_* = \inf\left\{\sum_i \sigma_i \,:\, W = \sum_i \sigma_i b_i a_i^T,\ \|b_i\|_2 = \|a_i\|_2 = 1, \sigma_i \geq 0\right\}.$$

The 1-nuclear norms measure costs of decomposition with respect to rank one elementary elements. This can be generalized to arbitrary elementary elements through *atomic norms* (Chandrasekaran et al., 2012). Consider a centrally symmetric set $\mathcal{A} \subset \mathbb{R}^d$, the *atom set.*

Definition 4. *The atomic norm (with atom set $\mathcal{A}$) of W is*

$$\|W\|_{\mathcal{A}} = \inf\left\{\sum_i \sigma_i \,:\, W = \sum_i \sigma_i a_i, a_i \in \mathcal{A}, \sigma_i \geq 0\right\}.$$

When $\text{conv}\,\mathcal{A}$ contains an open set that contains 0, this is indeed a norm over $\mathbb{R}^d$. Unlike nuclear norms, atomic norms include norms over vectors.

Exemple 12 (ℓ_1-norm). *The ℓ_1-norm on $\mathbb{R}^d$ is an atomic norm, with atom set*

$$\mathcal{A} = \{\pm e_i \,:\, i \in [d]\},$$

where e_i is the i-th vector from the canonical basis of $\mathbb{R}^d$.

2.3.4 The Max Norm

Recall that the trace norm can be obtained as the nuclear norm for $\|\cdot\|_A^\star = \|\cdot\|_B = \|.\|_2$. Hence, the trace norm measures the cost of decomposing a linear operator into rank one elementary elements of unit-Euclidean length. An interesting variation consists in using rank one elements with bounded coefficients. This can be obtained with $\|\cdot\|_A^\star = \|\cdot\|_B = \|.\|_\infty$. The corresponding nuclear norm is

$$\|W\|_{\infty\to 1}^\star = \inf\left\{\sum_i \sigma_i \,:\, W = \sum_i \sigma_i b_i a_i^T,\ \|b_i\|_\infty = \|a_i\|_\infty = 1, \sigma_i \geq 0\right\}.$$

For $d_1 = d_2 = 2$, one simply has $\|W\|_{\infty\to 1}^\star = \|W\|_\infty$. The associated unit ball is

$$\mathbf{B}_{\infty\to 1}^\star = \mathrm{conv}\left\{yf^T \,:\, \|y\|_\infty = \|f\|_\infty = 1\right\} = \mathrm{conv}\left\{yf^T \,:\, y_i, f_i \in \{-1,1\}\right\}.$$

This norm can be seen to provide a convex relaxation to the rank over the ℓ_∞-norm unit ball (Lee et al., 2008):

$$\left[\|W\|_{\infty\to 1}^\star\right]^2 \leq \mathrm{rank}(W)\,\|W\|_\infty.$$

Is this a practical choice of penalty? This norm is actually a constant factor approximation to the so-called *CUT norm*, and both can be shown to be NP-hard. Indeed, they both can be used to compute maximum cut values in graphs (Alon and Naor, 2006) by computing the norm of a well-chosen matrix. However, $\|.\|_{\infty\to 1}^\star$ is closely approximated by the *max norm* (or γ_2-norm) (Srebro, 2004).

Definition 5. *The max norm of $W \in \mathbb{R}^{d_1\times d_2}$ is defined equivalently as*

$$\|W\|_{max} = \inf_{UV^T=W} \|U\|_{2\to\infty}\|V\|_{2\to\infty} = \frac{1}{2}\inf_{UV^T=W}\left[\|U\|_{2\to\infty}^2 + \|V\|_{2\to\infty}^2\right].$$

In this expression, the number of columns of U and V are left unbounded. This is similar to the expression given for the trace norm, and is an example of a *factorization constant*. Here, instead of the average ℓ_2-norm of the rows of the factors, the maximum ℓ_2-norm of the rows of the factor are used:

$$\|U\|_{2\to\infty} = \max_{i\in[d_1]} \|U_{i,\cdot}\|_2.$$

This norm has also be referred to as the *Hadmard product operator norm* (Lee et al., 2008), due to the alternative expression

$$\|W\|_{\max} = \max_Q \frac{\|Q\circ W\|_{\mathrm{op}}}{\|Q\|_{\mathrm{op}}} = \max_{\|u\|_2=\|v\|_2=1} \|uv^T \circ W\|_*.$$

This latter expression has been used to derive sub differentials related to the max norm by Jalali and Srebro (2012). The max norm is not a spectral norm: unlike the trace norm, it cannot be expressed as a function of singular values only. However, unlike the previous cut-related norms, the max norm can be expressed as the optimal value of a semi-definite program, and thus can lead to tractable optimization problems:

$$\|W\|_{\max} = \inf\left\{t \,:\, \begin{matrix}\|\mathrm{diag}(A)\|_\infty \\ \|\mathrm{diag}(B)\|_\infty\end{matrix} \leq t \quad \text{and} \quad \begin{pmatrix} A & X \\ X^T & B\end{pmatrix} \succeq 0\right\},$$

over $A \in \mathbb{R}^{d_1 \times d_1}, B \in \mathbb{R}^{d_2 \times d_2}, t \in \mathbb{R}$. This property is called *semi-definite representability*. The max norm has numerous more properties, see previous references and references therein.

The aforementioned relation to $\|\cdot\|^\star_{\infty\to 1}$ can be seen as follows. Duality and Grothendick's inequality imply with $1.67 < K_G < 1.79$ that

$$K_G \|A\|^\star_{\infty\to 1} \leq \|A\|_{\max} \leq \|A\|^\star_{\infty\to 1}.$$

As with $\|\cdot\|^\star_{\infty\to 1}$, the max norm provides a convex relaxation to the rank over the ℓ_∞-norm, albeit less tight:

$$[\|W\|_{\max}]^2 \leq \left[\|W\|^\star_{\infty\to 1}\right]^2 \leq \operatorname{rank}(W)\|W\|_\infty.$$

The theoretical performances of the max norm have been analyzed in various settings, such as in clustering (Jalali and Srebro, 2012), or in matrix completion. In this last setting, Foygel and Srebro (2011) showed that the max norm is much less sensitive than the trace norm to the marginals of the sampling distributions: while the trace norm can be interpreted as penalizing the average row-norm of factorizations, the max norm penalizes the maximum row-norm. The local max norm $\|\cdot\|_{(\mathcal{R},\mathcal{C})}$ can be used to interpolate between the trace norm and the max norm through a proper choice of $\mathcal{R}$ and $\mathcal{C}$ (Foygel et al., 2012).

2.3.5 Other Heuristics and Open Problems

The experimenter may be faced with the problem of estimating not matrices, but higher dimensional objects such as tensors. Unlike with matrices, even symmetric tensors do not necessarily admit an orthogonal decomposition in terms of rank one atoms. Rank and equivalent of the trace norm have still be proposed, although these definitions are less canonical than in the matrix case. A standard idea consists in generalizing concepts from the matrix case, such as power iterations and Rayleigh quotients (Anandkumar et al., 2012). Another path to define convex relaxations consists in *unfolding*: given an order three tensor $T \in \mathbb{R}^{d_1 \times d_2 \times d_3}$, the unfolding along the k-th mode for $k \in [3]$ is a matrix rearrangement of the coefficients $T_{(k)} \in \mathbb{R}^{d_k \times \prod_{i \neq k} d_i}$. Tensor trace norm equivalents have been studied, where trace norms of the unfoldings are combined in various ways (Liu et al., 2013). Recently, other relaxations have been proposed by Romera-Paredes and Pontil (2013).

Finally, there are also some open problems. The trace norm is the convex envelope of the rank over the operator norm unit ball. However, in many settings (e.g., matrix completion with bounded ratings), imposing such a constraint on the maximum eigenvalue of matrices may not be relevant. As we have seen, the max norm and its nuclear norm approximation are convex relaxations to the rank over the ℓ_∞-norm ball. However, we do not know whether this nuclear norm is the tightest convex relaxation. This prompts the following question.

Open Problem 1. *What is the convex envelope of the rank over the ℓ_∞-norm unit ball? Formally, this amounts to determining*

$$\operatorname{conv}\left\{W \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(W) \leq r, \|W\|_\infty \leq 1\right\}.$$

In addition, is this homogeneous in r as with the trace norm?

This requires to understand relationships between a spectral-oriented quantity, and a coefficient-oriented quantity, which appears difficult. This motivates to consider different concepts of rank, or of complexity based on atoms.

Definition 6. *Let $\mathcal{A}$ be an atom set, and let $\Sigma_{\mathcal{A}}(W)$ denote the set of all optimal cost representations (σ_i, a_i) of W with respect to the corresponding atomic norm. The atomic representation length of W is*

$$\mu_{\mathcal{A}}(W) = \inf\{\|\sigma\|_0 : \sigma \in \Sigma_{\mathcal{A}}(W)\}.$$

The atomic sup norm of w is

$$M_{\mathcal{A}}(w) = \sup\{\|\sigma\|_\infty : \sigma \in \Sigma_{\mathcal{A}}(W)\}.$$

For $\mathcal{A} = \{\pm e_i : i \in [d]\}$, the atomic norm is the ℓ_1-norm, while $\mu_{\mathcal{A}}$ is the ℓ_0-norm, and $M_{\mathcal{A}}$ is the ℓ_∞-norm. Similarly, for $\mathcal{A} = \{uv^T : \|u\|_2 = \|v\|_2 = 1\}$, the atomic norm is the trace norm, while $\mu_{\mathcal{A}}$ is the rank, and $M_{\mathcal{A}}$ is the operator norm. In both examples above, optimal representations for $\mu_{\mathcal{A}}$ and $M_{\mathcal{A}}$ are obtained with atoms that are orthogonal to each others. An interesting question lies in how this can be generalized to non-orthogonal settings.

Proposition 5. *Let $\mathcal{A}$ be a centrally symmetric atom set. Then, $\mu_{\mathcal{A}}$ is sub-additive, and $\mu_{\mathcal{A}}(\lambda\cdot) = \mu_{\mathcal{A}}(\cdot)$ for any $\lambda \neq 0$. We also have*

$$M_A(w) \leq \|w\|_{\mathcal{A}} \leq \mu_{\mathcal{A}}(w) M_{\mathcal{A}}(w).$$

In addition, $M_{\mathcal{A}}(\lambda\cdot) = |\lambda| M_{\mathcal{A}}(\cdot)$ for any $\lambda \in \mathbb{R}$, and $M_{\mathcal{A}}(w) = 0$ if and only if $w = 0$.

Hence, $\|\cdot\|_{\mathcal{A}}$ is a convex relaxation to $\mu_{\mathcal{A}}$ over $\{w : M_{\mathcal{A}}(w) \leq 1\}$. This suggests the following problems.

Open Problem 2. *Under what conditions on $\mathcal{A}$ is $M_{\mathcal{A}}$ convex, and hence, a norm? Under what conditions is the atomic norm $\|\cdot\|_{\mathcal{A}}$ the tightest convex relaxation to $\mu_{\mathcal{A}}$ over $\{w : M_{\mathcal{A}}(w) \leq 1\}$?*

2.4 Measuring Quality of Regularizers and Theoretical Results

In the previous sections, we have presented many interesting regularizers to enforce variations of the notion of sparsity or low rank. A given penalty will be useful if it allows to guarantee a good generalization performance, or estimate the model with a minimum amount of training data, while remaining computationally tractable (so that the estimator can actually be computed in practice). There is a very large literature on theoretical analysis of the generalization performance of various penalized ERM programs (Bartlett and Mendelson, 2003; Shorack and Wellner, 2009), and we do not review this question in depth here. We will provide such an analysis in Chapter 4 in the context of multiple kernel learning. Instead, we focus in this section on theoretical results for the estimation problem.

2.4.1 Estimation: Exact and Robust Recovery

We will assume that there is a true linear model of the form (2.2) (vector form) or (2.4) (matrix form) to estimate. The usefulness of a given penalty in recovering a model of course depends a lot on how simple the true model is in the view of this penalty. Results that involve the complexity of the true object according to the penalty, e.g., $\Omega(\omega^\star)$ or similar geometric characterizations usually fall within the realm of *oracle inequalities*, and we give such an example in Chapter 3. On the other hand, results may involve the complexity of the true object according to a hard notion to which the penalty is a relaxation, e.g., $\|\omega^\star\|_0$. This is usually the ultimate objective, and we focus on such a type of results in this section. In the absence of noise, the problem is *exact recovery*: can we perfectly estimate $\omega^\star$ in the absence of noise? Formally, this consists in determining when

$$w^\star = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \Omega(w) \text{ s.t. } \langle x_i, w \rangle = y_i \text{ for } i \in [N]. \tag{2.5}$$

When there is noise, the problem is *robust recovery*: how well can we estimate in the presence of noise? Formally, this consists in determining upper bounds on $\|\widehat{w} - w^\star\|$, where

$$\widehat{w} = \operatorname*{arg\,min}_{w \in \mathbb{R}^d} \Omega(w) \text{ s.t. } \|Xw - y\| \leq \varepsilon \tag{2.6}$$

or variations on this, where $\|\cdot\|$ is usually the Euclidean norm. Note that these two formulations differ slightly from our penalized estimation problem, but this allows for more flexibility in order to describe common theoretical results.

The types of results that can be obtained depends greatly on the design. We consider Gaussian designs, and use the notation $A \in \mathbb{R}^{N \times d}$ for such designs: in this case, all entries of A are independent standard normals. The choice of Gaussian designs has a longstanding history. In particular, random Gaussian vectors are isotropic, and thus have an important connection to uniformly random directions. Most importantly, random projections onto Gaussian direction have been known for long to approximately preserve the Euclidean length, and pairwise Euclidean distances of points (Bourgain, 1985; Dasgupta and Gupta, 2003).

Proposition 6. *Let $x \in \mathbb{R}^d$, $0 < \varepsilon < 1$, and consider a random Gaussian design $A \in \mathbb{R}^{N \times d}$. Then, with probability at least $1 - 2e^{-(\varepsilon^2 - \varepsilon^3)\frac{N}{4}}$,*

$$(1 - \varepsilon)\|x\|_2^2 \leq \left\| \frac{1}{\sqrt{N}} Ax \right\|_2^2 \leq (1 + \varepsilon)\|x\|_2^2.$$

A major advance in recent years was to show that much more structure (such as the support of sparse vectors, or the singular spaces of low rank matrices) could also be recovered from Gaussian random projections, and most of the results we will discuss in the rest of this chapter are of this sort.

In problems of the form (2.5) or (2.6), the recovery is dictated by the geometry of the kernel of the design, that we refer to by $\ker A$. Due to rotational invariance of the Gaussian distribution, the subspace $\ker A$ is actually a random subspace of codimension N distributed uniformly at random.

2.4.2 High-Dimensional Convex Sets and Gaussian Width

Millman has proposed a pictorial representation of high-dimensional convex sets (Milman, 1998) that we reproduce in Figure 2.7: a convex set is made of a bulk which contains most of the mass, and of tentacles. Due to the fact that the representation is only two-dimensional, this includes hyperbolic boundaries. Before we consider the case of penalized estimation,

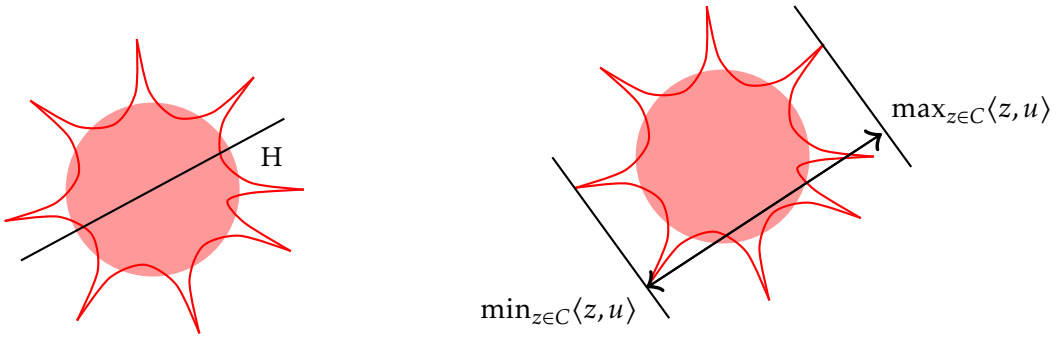

Figure 2.7: Pictorial representation of convex sets in high-dimension: intersection with a random subspace H (left), Gaussian width (right)

consider the following simple *feasibility problem*, where we simply take as estimator any point consistent with the data, and in the unit-ball of the regularizer:

$$\text{find } w \text{ s.t. } \Omega(w) \leq 1, Aw = y.$$

How accurately can we hope to estimate $w^\star$ using this type of program? This is fully characterized by the size of the intersection of the kernel of A and of $w^\star + \mathbf{B}_\Omega$. In particular, when this intersection is reduced to zero, exact recovery succeeds. The following notion of average width of a set will be useful to characterize when this is the case.

Definition 7. *The Gaussian width of $C \subset \mathbb{R}^d$ is*

$$w(C) = \mathbb{E}\left[\sup_{z\in C}\langle z, g\rangle\right],$$

where g is a standard Gaussian vector in $\mathbb{R}^d$.

This is a classical quantity in Gaussian process theory, and a treatment of Gaussian widths in our context can be found in (Chandrasekaran et al., 2012). The following are easy but useful properties of the Gaussian width. Here, the mean Euclidean length of a k-dimensional standard Gaussian vector is denoted by $\lambda_k \in [k/\sqrt{k+1}, \sqrt{k}]$.

Proposition 7. *The Gaussian width has the following properties: for any $C \subset \mathbb{R}^d$,*

- **Mean width formula.**

$$w(C) = \frac{\lambda_d}{2}\int_{\mathbf{S}_2}\left[\max_{z\in C}\langle z,u\rangle - \min_{z\in C}\langle z,u\rangle\right]du.$$

- **Convex hull.** $w(\mathrm{conv}(C)) = w(C)$,
- **Symmetrization.** $w(C) \leq w(C - C) \leq 2w(C)$,
- w *is invariant by orthogonal transformations or translations of* C,
- $w(C) = \sqrt{k}$ *if* C *is a linear subspace with dimension* k,
- *when* C *is the unit-ball of a norm* $\|\cdot\|$, *then* $w(C) = \mathbb{E}[\|g\|^\star]$.

The first property shows that the Gaussian width can be interpreted as the *mean width* in uniformly random directions, as illustrated in Figure 2.7. A central intuition here is that a random subspace H will tend to miss the tentacles of the convex set. We have the following result (Milman, 1985; Pajor and Tomczak-Jaegermann, 1986).

Proposition 8. *Let* H *be a random subspace of codimension* N, *and* C *a symmetric convex set, then, with probability at least* $1 - e^{-N}$, *for some constant* c_0,

$$\mathrm{diam}(C \cap H) \leq c_0 \frac{w(C)}{\sqrt{N}},$$

where $\mathrm{diam}(X)$ *is the diameter of set* X *in Euclidean norm.*

Although a bit longer to state formally, there exists similar matching lower bounds for the diameter of the intersection: see, for instance, (Giannopoulos et al., 2005), where such a lower bound is actually shown to hold for *all* subspaces of codimension N. As a consequence, setting $H = \ker A$, recovery up to ε-accuracy (in ℓ_2-norm) is possible for

$$N \geq \frac{w(C)^2}{\varepsilon^2}.$$

This approach allows to characterize when we can obtain ε-approximation using *feasibility problems*. To consider penalized estimation instead of simply feasibility, we need to look at slightly different geometrical objects, although similar results and intuitions will remain valid.

2.4.3 Optimality Conditions for Penalized Estimation

Consider

$$T_\Omega(w) = \mathrm{cone}\{z - w : \Omega(z) \leq \Omega(w)\},$$

the tangent cone at w to the corresponding scaled unit ball $\Omega(w)\mathbf{B}_\Omega = \{x \in \mathbb{R}^d : \Omega(x) \leq \Omega(w)\}$. Here, cone denotes the conical hull operator. The success of a noiseless penalized program is characterized by whether the nullspace of A intersects with the tangent cone. The following result can be found in (Chandrasekaran et al., 2012).

Proposition 9. *Let* $\Omega : \mathbb{R}^d \to \mathbb{R}_+$ *be a norm. In the noiseless setting,* $w^\star$ *is the unique optimal point of* (2.5) *if and only if*

$$\ker(A) \cap T_\Omega(w^\star) = \{0\}.$$

In the noisy setting, assume that noise is bounded as $\|\varepsilon\|_2 \leq \delta$, *and that for* $\nu \in]0,1[$, *it holds that*

$$\forall w \in T_\Omega(w^\star), \|Aw\|_2 \geq \nu \|w\|_2.$$

Then, all solutions to (2.6) *satisfy*

$$\|w - w^\star\|_2 \leq \frac{2\delta}{\nu}.$$

For robust recovery, the error bound on w is always larger than 2δ, which can be interpreted as the total standard deviation of the noise. The condition for exact recovery is illustrated in Figure 2.8.

The following result is a variation on Proposition 8 due to Gordon (1988), that is expressed directly in terms of the kernel of A, instead of an abstract random subspace H.

Proposition 10 (Escape Through a Mesh Phenomenon). *Let* C *be a closed set in* $\mathbb{R}^d$. *Then, for* $A \in \mathbb{R}^{N \times d}$ *a random Gaussian design,*

$$\mathbb{E}\left[\min_{w \in C \cap \mathbf{S}_2} \|Aw\|_2\right] \geq \lambda_N - w\left(C \cap \mathbf{S}_2\right).$$

In addition, $A \mapsto \min_{w \in C \cap \mathbf{S}_2} \|Aw\|_2$ *is Lipschitz with constant one with respect to the Frobenius norm.*

Intuitively, if the number N of measures is large enough compared to the squared Gaussian width, the lower bound is bounded away from zero such that elements in $C \cap \mathbf{S}_2$ cannot be in the kernel with high probability. Equipped with the conditions for recovery of Proposition 9, we can apply Proposition 10 to $T_\Omega(w^\star) \cap \mathbf{S}_2$. This can be used to deduce sufficient condition for exact or robust recovery with penalized estimation.

Proposition 11. *In the noiseless setting, exact recovery succeeds with high probability if* $N \geq w(T_\Omega(w^\star) \cap \mathbf{S}_2)^2 + 1$.

This result can actually be recovered using different tools from integral convex geometry, as we will see in the next section. The previous result emphasizes the importance of the

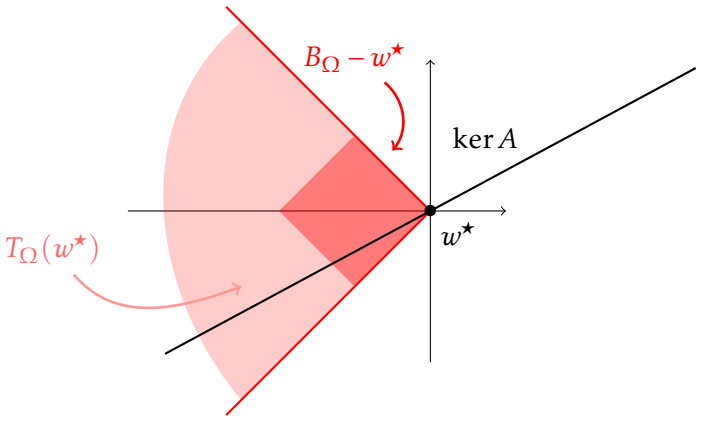

Figure 2.8: Illustration of optimality conditions for exact recovery for the ℓ_1-norm

geometry of the unit ball, and an interesting analysis for a given regularizer consists in identifying and classifying *special points* on the unit ball. In the following, we give a few examples of such points.

An *extremal point* of a convex set C is any point of C which cannot be decomposed as a convex combination of other points in C. Formally, a point x on the boundary of C is an extremal point of C if and only if $C \setminus \{x\}$ is still convex. The set of extremal points can be understood in the framework of atomic norms as an atom set of minimal size, but does not directly allow to characterize the smoothness of convex sets.

Definition 8. *Let C be a full-dimensional convex set. A boundary point x of C is said of order $\rho(x)$ if the intersection of all the supporting hyperplanes to C at x is an affine subspace of dimension $\rho(x)$.*

For a convex set in $\mathbb{R}^d$, a point $x \in C$ such that $\rho(x) = d - 1$ is said *smooth*. On the other hand, a point such that $\rho(x) = 0$ is referred to as a *vertex* (Gallier, 2008). The extremal points and the vertices do not necessarily coincide: a vertex is extremal, but the converse is not necessarily true (e.g., consider the Euclidean ball $\mathbf{B}_2$ for which the set of extremal points is the Euclidean sphere $\mathbf{S}_2$, but has no vertices). When $C = \mathbf{B}_\Omega$, points x where Ω is non-differentiable are such that $\rho(x) < d-1$ (i.e., they are not smooth), but not necessarily such that $\rho(x) = 0$ (i.e., they are not necessarily vertices).

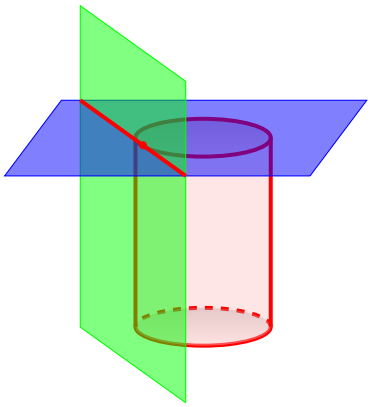

Figure 2.9: A boundary point of order 1 on a cylinder (red dot), and supporting hyperplanes (blue, green)

The points such that $0 < \rho(x) < d-1$ correspond to intermediate levels of smoothness. This is illustrated in Figure 2.9: we consider a convex set that is a cylinder (similarly to the unit ball for the trace norm over 2×2 symmetric matrices), and a point x on the boundary, indicated by a red dot. Unlike points in the interior of the top disk, this point is not smooth. However, this point is not a vertex either, as $\rho(x) = 1$. Indeed, there are two supporting hyperplanes to the cylinder at x, shown in blue and green, and their intersection is a line, shown in red. From similar considerations, one can show that the unit ball for the trace norm has no vertices, although any rank-deficient matrix is not a smooth point.

We note that even two norms related up to a constant factor can have different performances for recovery, as the geometry of their unit balls can be significantly different. This is, for instance, the case with the k-support norm and the elastic net: any point x such that

$\|x\|_0 < k$ is smooth for the ball of the k-support norm, but not for the ball of the elastic net. Similarly, the extremal points of the $\|\cdot\|^{\star}_{\infty\to 1}$-unit ball are all extremal points of the max norm unit ball, but the latter actually has strictly more (as otherwise, the two norms would coincide, contradicting lower bounds on Grothendick's constant K_G).

2.4.4 Kinematic Formula and Statistical Dimension

When working with convex programs and their optimality conditions, the main objects involved are convex cones. The kinematic formula for cones allows to quantify the probability that a convex cone C and another random rotated cone K intersect. Denote by Π_C the projection operator onto the convex set C. For a k-dimensional linear subspace L of $\mathbb{R}^d$, we also denote the probability that a random point on the sphere has a nontrivial projection on L by $I^d_k(\varepsilon) = \mathbb{P}(\|\Pi_L(\theta)\|_2^2 \geq \varepsilon)$. In order to generalize this to cones, consider the following quantities.

Proposition 12 (Conic Instrinsic Volumes)**.** *Let C be a closed convex cone in $\mathbb{R}^d$, then for any $\varepsilon \in [0,1]$ and θ uniformly distributed on the Euclidean unit sphere $\mathbf{S}_2$, there exists $d+1$ scalars*

$$(v_0(C),\dots,v_d(C))$$

such that

$$\mathbb{P}(\|\Pi_C(\theta)\|_2^2 \geq \varepsilon) = \sum_{k=0}^{d} v_k(C) I^d_k(\varepsilon).$$

The elements $v_0(C),\dots,v_d(C)$ are the intrinsic volumes of C.

The Kinematic Formula shows that the intrinsic volumes characterize the probability of intersection of two cones when one of them is rotated uniformly at random.

Theorem 1 (Cone Kinematic Formula)**.** *Assume that C and K are closed convex cones in $\mathbb{R}^d$, one of which is not a linear subspace, and denote by R a rotation uniformly at random. Then,*

$$\mathbb{P}\left(C \cap RK \neq \{0\}\right) = 2h_{d+1}(C \times K)$$

where

$$h_{d+1}(C\times K) = v_{d+1}(C\times K) + v_{d+3}(C\times K) + \dots$$

is the $(d+1)$-th half-tail functional.

In practice, the intrinsic volumes may be difficult to compute. Remarkably, Amelunxen et al. (2013) show that viewed as a discrete distribution, the sequence of intrinsic volumes concentrates around its mean, and that this can be used to provide an approximate version of the Kinematic formula. Although less precise, this approximation still allows to show a sharp threshold effect on the number of observations in terms of only a *single* dimension-like quantity of each of the cones. The mean intrinsic volume of a cone C is referred to as the *statistical dimension*, and can be expressed as

$$\Delta(C) = \sum_{k=0}^{d} k v_k(C) = \mathbb{E}\left[\|\Pi_C(g)\|_2^2\right]$$

for a closed convex cone C in $\mathbb{R}^d$, and $g \in \mathbb{R}^d$ a standard Gaussian vector. The characterization of the statistical dimension in terms of average norm of the projection of Gaussian vectors is paramount in allowing for practical calculations of the dimension. Notably, this is quite similar to the Gaussian width, and Amelunxen et al. (2013) actually show that, albeit arising from different considerations, these quantities are vey closely related, as for a cone C,

$$w^2(C \cap \mathbf{S}_2) \leq \Delta(C) \leq w^2(C \cap \mathbf{S}_2) + 1.$$

This leads to the following result (Amelunxen et al., 2013).

Theorem 1 (Approximate Kinematic Formuma for Inverse Problems). *Let $\eta \in]0,1[$ and Ω be a norm on $\mathbb{R}^d$. For $A \in \mathbb{R}^{N \times d}$ with independent standard Gaussian entries, it holds that*

$$N \leq \Delta(T_\Omega(w^\star)) - \alpha_\eta \sqrt{d} \Rightarrow \textit{exact recovery with probability less than } \eta,$$
$$N \geq \Delta(T_\Omega(w^\star)) + \alpha_\eta \sqrt{d} \Rightarrow \textit{exact recovery with probability at least } 1 - \eta,$$

with $\alpha_\eta = \sqrt{8 \log(4/\eta)}$.

This completes many previous results which provided either only upper or lower bounds. This is a remarkable result, which shows the existence of a sharp threshold on the number of observations for exact recovery. Although the first results of this type were in the setting of ℓ_1-norm penalized estimation, this result shows that this threshold phenomenon actually applies to a wide class of random convex programs.

2.4.5 Examples

We now give examples of regularizers, together with their geometric characteristics and statistical dimensions. The unit balls corresponding to some of these norms are shown in Figure 2.10. We indicate asymptotic orders of magnitude without constants using the tilde symbol $\sim$.

Figure 2.10: Unit balls for ℓ_∞-norm, ℓ_2-norm, ℓ_1-norm, elastic-net norm ($\gamma = 1$), 2-support norm

Unit balls. Although not directly useful in characterizing sample complexity, we give the widths of a few classical unit balls. Using the characterization of the Gaussian width as the expected dual norm of a standard Gaussian vector, we easily obtain the following (Vershynin, 2011):

$$w(\mathbf{B}_1)^2 \sim \log d, \quad w(\mathbf{B}_2)^2 \sim d, \quad w(\mathbf{B}_\infty) \sim d^2.$$

Sparse vectors with the ℓ_1-norm. Consider the set of k-sparse vectors

$$S_k = \left\{ w \in \mathbb{R}^d \,:\, \|w\|_0 \leq k \right\},$$

such that $w(S_k \cap \mathbf{B}_2)^2 \sim k \log(2d/k)$ (Plan and Vershynin, 2013b). In addition, let $w^\star \in S_k$, and $\Omega = \|\cdot\|_1$, then

$$w(T_\Omega(w^\star) \cap \mathbf{S}_2)^2 \leq 2k \log(d/k) + 5/4k.$$

This upper bound can be found in (Chandrasekaran et al., 2012). This is a key result of sparsity: the sample complexity only depends logarithmically on the ambient dimension. This logarithmic factor of order $\log \binom{d}{k}$ corresponds to the cost of not knowing what the support of the model is.

Sparse vectors with the s-support norm. For Ω equal to the s-support norm (for s that does not necessarily match k), one has similarly

$$w(T_\Omega(w^\star) \cap \mathbf{S}_2)^2 \leq \left[\sqrt{2s \log(d+2)} + \sqrt{s} \right]^2 \left\lceil \frac{k}{s} \right\rceil + k.$$

This upper bound is from Chatterjee et al. (2014), and is based on a technique proposed for the analysis of the Group Lasso penalty (Rao et al., 2012). This recovers the extreme cases of $\Omega = \|\cdot\|_1$ (when $s = 1$) and $\Omega = \|\cdot\|_2$ (when $s = d$).

Sign vectors. Consider the set of k-saturated vectors

$$C_k = \left\{ w \in \mathbb{R}^d \,:\, \operatorname{card}\{i \,:\, |w_i| = \|w\|_\infty\} = k \right\},$$

such that $w(C_k \cap \mathbf{B}_2)^2 = d^2$. Note that $C_k \cap \mathbf{B}_\infty$ is actually a k-face of the unit ball for the ℓ_∞-norm. In addition, let $w^\star \in C_k$, and $\Omega = \|\cdot\|_\infty$, then

$$\Delta(T_\Omega(w^\star)) = d - \frac{s-k}{2}.$$

The derivation of the exact value of the statistical dimension can be found in (Amelunxen et al., 2013).

Low-rank matrices with the trace norm. Consider the set of rank r matrices

$$L_r = \left\{ W \in \mathbb{R}^{d_1 \times d_2} \,:\, \operatorname{rank}(W) = r \right\}.$$

For $W^\star \in L_r$, and $\Omega = \|\cdot\|_*$, then

$$w(T_\Omega(W^\star) \cap \mathbf{S}_2)^2 \leq 3r(d_1 + d_2 - r).$$

This result can be found in (Chandrasekaran et al., 2012). Analyses of penalized estimation with the max norm usually use other routes, see, for instance, (Srebro and Shraibman, 2005; Cai and Zhou, 2013) for results in matrix completion based on Rademacher complexities.

Nonnegative entries. In most of the previous cases, one may additionally incorporate the knowledge that the model only has nonnegative entries, such that the ambient space becomes $\mathbb{R}^d_+$ (Donoho and Tanner, 2010). In some particular settings (e.g., using non-Gaussian designs), this has been shown to lead to a surprising sample complexity of order k for exact recovery of k-sparse nonnegative vectors (Donoho and Tanner, 2005; Wang and Tang, 2009) using a simple feasibility problem as in Section 2.4.2. This corresponds to strictly more than a constant factor improvement over sample complexities of order $k \log d$, as this is completely dimension-independent. In other settings, however, it is frequent to observe only constant factor improvements.

2.4.6 Estimation with Non-gaussian Designs

Drawing inspiration from the domain of random projections at large, various conditions have been proposed (usually in specific cases) to ensure that a non-Gaussian (but possibly still random) design may lead to exact recovery. This includes the Random Fourier ensemble, the Bernoulli ensemble or Count-Min matrices (Berinde et al., 2008). Many of the associated analyses are specific to ℓ_1-norm minimization (and ℓ_0-sparse models), and rely on properties such as Restricted Isometry Property (Berinde et al., 2008), or the Null Space Property (Cohen et al., 2009). Testing whether these properties hold on fixed matrices is usually NP-hard, although relaxations have been proposed (d'Aspremont and El Ghaoui, 2011). Recently, an imposed real world design has been shown to exhibit a phase transition (Vattikuti et al., 2014), but such examples remain scarce.

The use of nonlinear models and designs is of course an interesting area (Plan and Vershynin, 2013b). An interesting such example consists in using quantized measurements. Indeed, in all the previous designs, measurements are assumed to be a vector in $\mathbb{R}^N$, where components are encoded with infinite precision. In practice, this is rarely the case as most measurement systems have a limited dynamic and can handle a limited number of bits per measurement. In addition to this, the floating point arithmetic used in the computers has also a limited bit resolution. Hence, the bit resolution is potentially limited at many stages. Consider a quantization operator $Q : \mathbb{R} \to \{1, \dots, K\}$ over an alphabet $(\mathfrak{c}_1, \dots, \mathfrak{c}_K) \in \mathbb{R}^K$ of length K. A quantized linear model can be obtained as

$$y_i = Q(\langle a_i, w \rangle), i \in [N].$$

In the extreme case, the alphabet can reduce to $(-1, 1)$ only: that is, only the sign of the projection is measured:

$$y_i = \operatorname{sign}(\langle a_i, w \rangle), i \in [N].$$

This model is referred to as *one-bit* compressed sensing. This also corresponds to random hyperplane hashing, which is a hashing scheme for approximating the angular similarity

$$\theta(w_1, w_2) = 1 - \frac{1}{\pi} \cos^{-1}\left(\frac{\langle w_1, w_2 \rangle}{\|w_1\|_2 \|w_2\|_2}\right)$$

between vectors $w_1, w_2 \in \mathbb{R}^d$. Consider the Hamming distance $d_H(x, y) = \sum_{i=1}^N x_i y_i$ over the boolean hypercube $\{0, 1\}^N$. An estimation of the angle between any two pair of vectors can be obtained through the Hamming distance between the one-bit observations in the corresponding models: with N observations, and A a random Gaussian design,

$$\mathbb{E}\left[\frac{1}{N} d_H(\operatorname{sign}(Aw_1), \operatorname{sign}(Aw_2))\right] = \frac{\theta(w_1, w_2)}{\pi}.$$

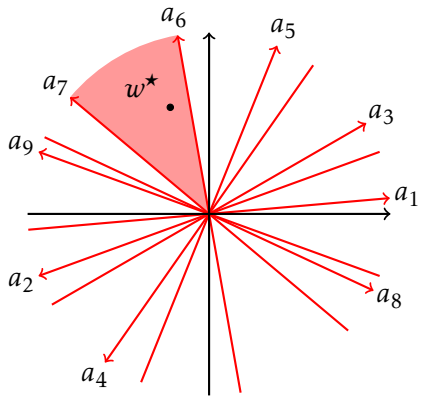

Figure 2.11: Illustration of random hyperplane hashing

This is natural, as illustrated in Figure 2.11, as the Hamming distance counts the number of circular sectors where both vectors get projected, and averaging over many hyperplanes provides an estimation of the angle. A similar intuition carries over to the estimation problem, and recently, many works have considered the question of characterizing when exact or robust recovery is possible with this type of measurements (Gupta et al., 2010). As shown by Plan and Vershynin (2013a), in the one-bit setting, robust recovery up to precision ε can be achieved with about the same number of measurements based on the squared Gaussian width, albeit with a worse dependence on ε.

3

Estimation of Sparse and Low Rank Matrices

Contents

3.1 Introduction . . . 64
3.1.1 Model . . . 65
3.1.2 Main Examples . . . 65
3.1.3 Outline . . . 65
3.1.4 Notation . . . 66
3.2 Oracle Inequality . . . 66
3.3 Generalization Error in Link Prediction . . . 67
3.4 Algorithms . . . 68
3.4.1 Proximal Operators . . . 68
3.4.2 Generalized Forward-backward Splitting . . . 69
3.4.3 Incremental Proximal Descent . . . 69
3.4.4 Positive Semi-definite Constraint . . . 70
3.5 Recovering Clusters . . . 70
3.6 Numerical Experiments . . . 71
3.6.1 Synthetic Data . . . 71
3.6.2 Real Data Sets . . . 71
3.7 Discussion . . . 73
3.7.1 Other Loss Functions . . . 73
3.7.2 Optimization . . . 74
3.7.3 Geometry . . . 74
3.7.4 Factorization Methods . . . 75
3.8 Proofs . . . 78
3.8.1 Sketch of Proof of Proposition 13 . . . 78
3.8.2 Proof of Proposition 16 . . . 79

3.1 Introduction

Matrix estimation is at the center of many modern applications and theoretical advances in the field of high dimensional statistics. The key element which differentiates this problem from standard high dimensional vector estimation lies in the structural assumptions that can be formulated in this context. Indeed, the notion of sparsity assumption has been transposed into the concept of low rank matrices and opened up the way to numerous achievements (Srebro, 2004; Cai et al., 2008). In this chapter, we argue that being low rank is not only an equivalent of sparsity for matrices but that being low rank and sparse can actually be seen as two orthogonal concepts. The underlying structure we have in mind is that of a block diagonal matrix. This situation occurs for instance in covariance matrix estimation in the case of groups of highly correlated variables or when denoising/clustering social graphs.

Efficient procedures developed in the context of sparse model estimation mostly rely on the use of ℓ_1-norm regularization (Tibshirani, 1996a). Natural extensions include cases where subsets of related variables are known to be active simultaneously (Yuan and Lin, 2006). These methods are readily adapted to matrix valued data and have been applied to covariance estimation (El Karoui, 2009; Bien and Tibshirani, 2010) and graphical model structure learning (Banerjee et al., 2008; Friedman et al., 2008). In the low rank matrix completion problem, the standard relaxation approach leads to the use of the trace norm as the main regularizer within the optimization procedures (Srebro et al., 2005; Koltchinskii et al., 2011a) and their resolution can either be obtained in closed form (loss measured in terms of Frobenius norm) or through iterative proximal solutions (Combettes and Pesquet, 2011; Beck and Teboulle, 2009) (for general classes of losses). However, solutions of low rank estimation problems are in general not sparse at all, while denoising and variable selection on matrix-valued data are blind to the global structure of the matrix and process each variable independently. In this chapter, we study the benefits of using the sum of ℓ_1 and trace-norms as regularizer. This sum of regularizers on the same object allows to benefit from the virtues of both of them, in the same way as the elastic-net (Zou and Hastie, 2005a) combines the sparsity-inducing property of the ℓ_1 norm with the smoothness of the quadratic regularizer.

The trace norm and the ℓ_1 regularizers have already been combined in a different context. In Robust PCA (Candès et al., 2011) and related literature, the signal W is assumed to have an additive decomposition $W = X + Y$ where X is sparse and Y low rank. Note that W is not in general sparse nor low rank and that this decomposition is subject to identifiability issues, as analyzed, e.g., in (Chandrasekaran et al., 2011). The decomposition is recovered by using ℓ_1-norm regularization over X and trace norm regularization over Y. This technique has been successfully applied to background substraction in image sequences, to graph clustering (Jalali et al., 2011) and covariance estimation (Luo, 2011). We would like to emphasize that this type of *demixing* problems are largely different from what we consider here.

Here, we consider the different situation where a single matrix W is both sparse and low rank at the same time. We demonstrate the applicability of our mixed penalty on different problems. We develop proximal methods to solve these convex optimization problems and we provide numerical evidence as well as theoretical arguments which illustrate the trade-off which can be achieved with the suggested method.

3.1.1 Model

For a matrix $W = (W_{i,j})_{i,j}$, we will consider the following matrix norms: $\|W\|_1 = \sum_{i,j} |W_{i,j}|$ and $\|W\|_* = \sum_{i=1}^{\text{rank}(W)} \sigma_i(W)$. We consider the following setup. Let $Y \in \mathbb{R}^{d\times d}$ be an observed matrix and $\ell : \mathbb{R}^{d\times d} \times \mathbb{R}^{d\times d} \to \mathbb{R}_+$ a loss function over matrices. We introduce the following optimization problem:

$$\underset{W\in\mathcal{W}}{\arg\min}\, [\ell(W,Y) + \gamma\|W\|_1 + \tau\|W\|_*]$$

for some convex admissible set $\mathcal{W} \subset \mathbb{R}^{d\times d}$ and nonnegative regularization parameters γ and τ.

3.1.2 Main Examples

The underlying assumption in this work is that the unknown matrix to be recovered has a block-diagonal structure. We now describe the main modeling choices through the following motivating examples:

- Example 5: *Covariance matrix estimation* - the matrix Y represents a noisy estimate of the true covariance matrix obtained for instance with very few observations; the search space is $\mathcal{W} = S_+$ the class of positive semi-definite matrices; for the loss, we consider the squared norm $\ell(W,Y) = \|W - Y\|_F^2$.

- *Graph denoising* - the matrix Y is the adjacency matrix of a noisy graph with both irrelevant and missing edges; the search space is all of $\mathcal{W} = \mathbb{R}^{d\times d}$ and the coefficients of a candidate matrix estimate S are interpreted as signed scores for adding/removing edges from the original matrix Y; again, we use $\ell(W,Y) = \|W - Y\|_F^2$.

- Example 6: *Link prediction* - the matrix Y is the adjacency matrix of a partially observed graph: entries are 0 for both not-existing and undiscovered links. The search space is unrestricted as before and the matrix W contains the scores for link prediction; the ideal loss function is the empirical average of the zero-one loss for each coefficient

$$\ell_E(W,Y) = \frac{1}{|E|} \sum_{(i,j)\in E} 1\{(Y_{i,j} - 1/2)\cdot W_{i,j} \le 0\},$$

  where E is the set of edges in Y. However, as in classification theory, practical algorithms should use a convex surrogate (e.g., the hinge loss).

3.1.3 Outline

The remainder of this chapter is organized as follows. Sections 3.2 and 3.3 are devoted to theoretical results on the interplay between sparse and low rank effects. Section 3.4 presents algorithms used for resolution of the optimization problem. Section 3.5 discusses how clusters can be recovered from matrix estimates and Section 3.6 is devoted to numerical experiments. We close this chapter by giving some perspectives in Sections 3.7. This chapter is an extended version of a paper with Emile Richard and Nicolas Vayatis which has appeared in the proceedings of ICML (Richard et al., 2012).

3.1.4 Notation

In the sequel, the projection of a matrix M onto a convex set $\mathcal{W}$ is denoted by $\Pi_{\mathcal{W}}(M)$. The matrix M_+ is the componentwise positive part of the matrix M, and $\mathrm{sgn}(M)$ is the sign matrix associated to M with the convention $\mathrm{sgn}(0) = 0$. The componentwise or Hadamard product of matrices is denoted by $\circ$. The class S_+ of matrices is the convex cone of positive semi-definite matrices in $\mathbb{R}^{d\times d}$. The sparsity index of M is $\|M\|_0 = \mathrm{card}\{M_{i,j} \neq 0\}$ and the Frobenius norm of a matrix M is defined by $\|M\|_F^2 = \sum_{i,j} M_{i,j}^2$. The operator norm of M is $\|M\|_{\mathrm{op}} = \sup_{x\,:\,\|x\|_2=1} \|Mx\|_2$, while $\|M\|_\infty = \max|M_{i,j}|$

3.2 Oracle Inequality

The next result shows how matrix recovery is governed by the trade-off between the rank and the sparsity index of the unknown target matrix, or by their convex surrogates: the trace norm and the ℓ_1-norm.

Proposition 13. *Let $W^\star \in \mathbb{R}^{d\times d}$ and $Y = W^\star + \epsilon$ with $\epsilon \in \mathbb{R}^{d\times d}$ having i.i.d. entries with zero mean. Assume for some $\alpha \in [0,1]$ that $\tau \geq 2\alpha\|\epsilon\|_{op}$ and $\gamma \geq 2(1-\alpha)\|\epsilon\|_\infty$. Let*

$$\widehat{W} = \operatorname*{arg\,min}_{W\in\mathcal{W}} \Big[\|W - Y\|_F^2 + \gamma\|W\|_1 + \tau\|W\|_*\Big].$$

Then,

$$\|\widehat{W} - W^\star\|_F^2 \leq \inf_{W\in\mathcal{W}} \Big[\|W - W^\star\|_F^2 + 2\gamma\|W\|_1 + 2\tau\|W\|_*\Big],$$

$$\|\widehat{W} - W^\star\|_F^2 \leq \Big[2\gamma\|W^\star\|_1 + 2\tau\|W^\star\|_*\Big] \wedge \left[\gamma\sqrt{\|W^\star\|_0} + \tau\sqrt{\mathrm{rank}(W^\star)}\,\frac{\sqrt{2}+1}{2}\right]^2.$$

The techniques used in the proof, that we defer to Section 3.8.1, are similar to those introduced in (Koltchinskii et al., 2011a). Note that the upper bound interpolates between the results known for trace-norm penalization and Lasso. In fact, for $\alpha = 0$, τ can be set to zero, and we get a sharp bound for Lasso, while the trace-norm regression bounds of (Koltchinskii et al., 2011a) are obtained for $\alpha = 1$.

From a theoretical point of view, Proposition 13 provides us with performance guarantees when the regularization parameters are large enough. From random matrix theory, the operator norm of a random Gaussian matrix is known to concentrate around $\sqrt{d}$, which enforces a stringent constraint on τ for $\tau \geq 2\alpha\|\epsilon\|_{\mathrm{op}}$ to hold with high probability. Similarly, the ∞-norm $\|\epsilon\|_\infty$ can be bounded by $\|\epsilon\|_{\mathrm{op}}$ or using the multivariate Tchebycheff inequality of Olkin and Pratt (1958) which implies that the condition $\gamma \geq 2(1-\alpha)\|\epsilon\|_\infty$ is satisfied with probability $1-\delta$ when $\gamma = \Omega\Big((1-\alpha)\frac{2d\sigma}{\delta}\Big)$. In practice, γ should not exceed the order of magnitude of the entries of the matrix, as this leads to a trivial zero solution. Asymptotically, to keep the sparsity regularization parameter γ of the order of magnitude of elements of the observation matrix Y, the free parameter α must be chosen so that $1 - \alpha_d \sim_d \frac{1}{d}$. This gives the same asymptotic behavior in $O(\sqrt{d})$ for the lower bound on τ as in matrix completion.

The proof can also easily be extended to more general fixed matrix designs where observations are of the form $y_i = \langle X_i, W^\star\rangle + \varepsilon_i$ for $i \in [N]$. In this case, the bound is in

terms of the design-dependant norm given by

$$\|\widehat{W} - W^\star\| = \sqrt{\sum_{i=1}^{N} \langle X_i, \widehat{W} - W^\star \rangle^2}.$$

Proposition 13 corresponds to a design $\{X_1, \ldots, X_{d^2}\}$ which is the canonical basis of $\mathbb{R}^{d\times d}$.

3.3 Generalization Error in Link Prediction

We dwell for a moment on the task of link prediction in order to illustrate how rank and sparsity constraints can help in this setting. Consider a graph over d vertices with adjacency matrix $Y \in \{0,1\}^{d\times d}$, and a subset E of coordinates $(i,j) \in [d] \times [d]$ of this matrix that have been observed: for $(i,j) \in E$, the value $Y_{i,j} \in \{0,1\}$ is known to the experimenter. We set out to predict the values of the remaining entries of Y by finding a sparse rank r predictor $W \in \mathbb{R}^{d\times d}$ with small zero-one loss

$$\ell(W,Y) = \frac{1}{d^2} \sum_{(i,j)\in[d]\times[d]} 1\{(Y_{i,j} - 1/2) \cdot W_{i,j} \leq 0\}$$

by minimizing the empirical zero-one loss

$$\ell_E(W,Y) = \frac{1}{|E|} \sum_{(i,j)\in E} 1\{(Y_{i,j} - 1/2) \cdot W_{i,j} \leq 0\}.$$

The objective of a generalization bound is to relate $\ell(W,Y)$ with $\ell_E(W,Y)$. In the case of the sole rank constraint, Srebro (2004) remarked that all low rank matrices with the same sign pattern are equivalent in terms of loss and applied a standard argument for generalization in classes of finite cardinality. In the work of Srebro, a beautiful argument is used to upper bound the number of distinct sign configurations for predictors of rank r

$$s_{\mathrm{lr}}(d,r) = \mathrm{card}\{\mathrm{sgn}(W) \,|\, W \in \mathbb{R}^{d\times d}, \mathrm{rank}(W) \leq r\}|$$

leading to the following generalization performance: for $\delta > 0$, $Y \in \{0,1\}^{d\times d}$ and with probability $1-\delta$ over choosing a subset E of entries in $\{1,\ldots,d\}^2$ uniformly among all subsets of $|E|$ entries, we have for any matrix W of rank at most r and $\Delta(d,r) = \left(\frac{8ed}{r}\right)^{2dr}$ that

$$\ell(W,Y) < \ell_E(W,Y) + \sqrt{\frac{\log\Delta(d,r) - \log\delta}{2|E|}}. \tag{3.1}$$

We consider the class of sparse rank r predictors

$$\mathcal{M}(d,r,s) = \{UV^T \,|\, U,V \in R^{d\times r}, \|U\|_0 + \|V\|_0 \leq s\}$$

and let $s_{\mathrm{splr}}(d,r,s)$ be the number of sign configurations for the set $\mathcal{M}(d,r,s)$. By upper bounding the number of sign configurations for a fixed sparsity pattern in (U,V) using an argument similar to (Srebro, 2004), a union bound gives

$$s_{\mathrm{splr}}(d,r,s) \leq \Gamma(d,r,s) = \left(\frac{16ed^2}{s}\right)^s \binom{2dr}{s}.$$

Using the same notations as previously, we deduce from this result the following generalization bound: with probability $1-\delta$ and for all $W \in \mathcal{M}(d,r,s)$,

$$\ell(W,Y) < \ell_E(W,Y) + \sqrt{\frac{\log \Gamma(d,r,s) - \log \delta}{2|E|}}. \tag{3.2}$$

In general, the square-root deviation term in bound (3.2) is smaller than the one in (3.1) for sufficiently large values of d as shown in the next proposition. The two bounds coincide when $s = 2dr$, that is, when (U,V) is dense and there is no sparsity constraint.

Proposition 14. *For $r_d = d\beta$ with $\beta \in]0,1]$ and $s_d = d\alpha$ with $\alpha \leq 2\beta$,*

$$\frac{\Delta(d,r_d)}{\Gamma(d,r_d,s_d)} = \Omega\left(\left[\frac{8ed(\beta d - \alpha)}{(\beta d)^2}\right]^{2d^2\beta}\right),$$

which diverges when d goes to infinity.

Proof. The result follows from the application of Stirling's formula. □

By considering a predictor class of lower complexity than low rank matrices, we can thus achieve better generalization performances. Although this illustrates that combining rank and sparsity allows to improve generalization performance, minimizing the classification error over $\mathcal{M}(d,r,s)$ does not lead to a practical procedure.

3.4 Algorithms

We now present how to solve the optimization problem with the proposed regularizer. We consider a loss function $\ell(W,Y)$ convex and differentiable in W, and assume that its gradient is Lipschitz with constant L and can be efficiently computed. This is, in particular, the case for the previously mentioned squared Frobenius norm, and for other classical choices such as the logistic loss.

3.4.1 Proximal Operators

We encode the presence of a constraint set $\mathcal{W}$ using the indicator function $\delta_{\mathcal{W}}(W)$ that is zero when $W \in \mathcal{W}$ and ∞ otherwise, leading to

$$\widehat{W} = \operatorname*{arg\,min}_{W \in \mathbb{R}^{d\times d}} \left[\ell(W,Y) + \tau\|W\|_1 + \gamma\|W\|_* + \delta_{\mathcal{W}}(W)\right].$$

This formulation involves a sum of a convex differentiable loss and of convex non-differentiable regularizers which renders the problem non trivial. A string of algorithms have been developed for the case where the optimal solution is easy to compute when each regularizer is considered in isolation. Formally, this corresponds to cases where the proximal operator defined for a convex regularizer $\Omega : \mathbb{R}^{d\times d} \to \mathbb{R}$ at a point Z by

$$\operatorname*{prox}_{\Omega}(Z) = \operatorname*{arg\,min}_{W \in \mathbb{R}^{d\times d}} \left[\frac{1}{2}\|W - Z\|_F^2 + \Omega(W)\right]$$

is easy to compute for each regularizer taken separately. See (Combettes and Pesquet, 2011; Parikh and Boyd, 2013) for a broad overview of proximal methods.

The proximal operator of the indicator function is simply the projection onto $\mathcal{W}$, which justifies the alternate denomination of generalized projection operator for prox_Ω. The proximal operator for the trace norm is given by the shrinkage operation as follows (Beck and Teboulle, 2009). If $Z = U \operatorname{diag}(\sigma_1, \cdots, \sigma_d) V^T$ is the singular value decomposition of Z,

$$\text{SHR}_\tau(Z) = \underset{\tau\|\cdot\|_*}{\text{prox}}(Z) = U \operatorname{diag}((\sigma_i - \tau)_+)_i V^T.$$

Similarly, the proximal operator for the ℓ_1-norm is the soft thresholding operator

$$\text{ST}_\gamma(Z) = \underset{\gamma\|\cdot\|_1}{\text{prox}} = \text{sgn}(Z) \circ (|Z| - \gamma)_+.$$

3.4.2 Generalized Forward-backward Splitting

The family of Forward-Backward splitting methods are iterative algorithms applicable when there is only one non-differentiable regularizer. These methods alternate a gradient step and a proximal step, leading to updates of the form

$$W_{k+1} = \underset{\theta\Omega}{\text{prox}}(W_k - \theta \operatorname{grad}_W \ell(W_k, Y)),$$

for a stepsize θ. In particular, this corresponds to projected gradient descent when Ω is the indicator function of a convex set. On the other hand, Douglas-Rachford splitting tackles the case of $q \geq 2$ terms but does not benefit from differentiability. A generalization of these two setups has been recently proposed in (Raguet et al., 2013) under the name of Generalized Forward-Backward, which we specialize to our problem in Algorithm 1. The proximal operators are applied in parallel, and the resulting (Z_1, Z_2, Z_3) is projected onto the constraint that $Z_1 = Z_2 = Z_3$ which is given by the mean. The auxiliary variable Z_3 can be simply dropped when $\mathcal{W} = \mathbb{R}^{d \times d}$. The algorithm converges under very mild conditions when the step size θ is smaller than $\frac{2}{L}$.

Algorithm 1 Generalized Forward-Backward

Initialize $W, Z_1, Z_2, Z_3 = Y$, q = 3
repeat
  Compute $G = \nabla_W \ell(W, Y)$.
  Compute $Z_1 = \text{prox}_{q\theta\tau\|\cdot\|_*}(2W - Z_1 - \theta G)$
  Compute $Z_2 = \text{prox}_{q\theta\gamma\|\cdot\|_1}(2W - Z_2 - \theta G)$
  Compute $Z_3 = \Pi_{\mathcal{W}}(2W - Z_3 - \theta G)$
  Set $W = \frac{1}{q}\sum_{k=1}^{q} Z_k$
until convergence
return W

3.4.3 Incremental Proximal Descent

Although Algorithm 1 performs well in practice, the $O(d^2)$ memory footprint with a large leading constant due to the parallel updates can be a drawback in some cases. As a consequence, we mention a matching serial algorithm (Algorithm 2) introduced in (Bertsekas,

2011) that has a flavor similar to multi-pass stochastic gradient descent. We present here a version where updates are performed according to a cyclic order, although random selection of the order of the updates is also possible.

Algorithm 2 Incremental Proximal Descent

Initialize $W = Y$
repeat
  Set $W = W - \theta \nabla_W \ell(W, Y)$
  Set $W = \text{prox}_{\theta\tau\|\cdot\|_*}(W)$
  Set $W = \text{prox}_{\theta\gamma\|\cdot\|_1}(W)$
  Set $W = \Pi_{\mathcal{W}}(W)$
until convergence
return W

3.4.4 Positive Semi-definite Constraint

For any positive semi-definite matrix, we have $\|Z\|_* = \text{Tr}(Z)$. The simple form of the trace norm allows to take into account the positive semi-definite constraint at no additional cost, as the shrinkage operation and the projection onto the convex cone of positive semi-definite matrices can be combined into a single operation.

Lemma 1. *For $\tau \geq 0$ and $W \in \mathbb{R}^{d\times d}$,*

$$\underset{\tau\|\cdot\|_*+\delta_{S_+}}{\text{prox}}(W) = \underset{Z\geq 0}{\arg\min}\left[\frac{1}{2}\|Z-W\|_F^2 + \tau\|Z\|_*\right] = \Pi_{S_+}(W-\tau I).$$

3.5 Recovering Clusters

The estimators described previously may allow to estimate a covariance or adjacency matrix, but not to directly recover clusters or groups. Although outside of the scope of this work, we mention here a few methods which could be used.

Consider a symmetric and sparse square matrix $W \in \mathbb{R}^{d\times d}$. This can be viewed as the adjacency matrix of a graph $G = (V, E)$, where $V = [d]$, and $E = \{(i,j) : W_{i,j} > 0\}$. The graph bandwidth allow to measure to what extent a graph corresponds to a chain of some order.

Definition 9 (Graph Bandwidth)**.** *The graph bandwidth of matrix W (or graph G) is*

$$GBW(W) = \min_{f\in\mathfrak{S}_d} \max\left\{|f(i) - f(j)| : W_{i,j} > 0\right\},$$

where $\mathfrak{S}_d$ is the set of all permutations $f : [d] \to [d]$.

For instance,

$$\text{GBW}\begin{bmatrix} 1 & & 1 \\ & \ddots & \\ 1 & & 1 \end{bmatrix} = \text{GBW}\begin{bmatrix} 1 & 1 & & \\ 1 & 1 & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} = 2.$$

The path graph on d vertices has bandwidth 1, while a star over $k+1$ vertices has bandwidth $\lfloor (k-1)/2 \rfloor + 1$. The Cuthil-McKee algorithm (Cuthill and McKee, 1969) seeks to simultaneously permute the lines and rows of W such that the resulting matrix is banded with the smallest bandwidth. This is done in a Breadth-First Search (BFS) manner: fix a vertex, and visit each of its neighbor in the reverse order of their degrees. This is an heuristic, as both the optimal ordering and the associated cost are NP-hard to find, but that can still be used to reorder estimators. More elaborate methods have been proposed to estimate the bandwidth (Blum et al., 1998; Dunagan and Vempala, 2001), although these cannot usually be used on large scale data. Alternate notions of bandwidths have been proposed, such as the 2-SUM, where the squared difference is used in place of the absolute difference. In this case as well, various relaxations have been proposed by Fogel et al. (2013).

Although this may allow for easier visual inspection and representation, this still does not output clusters. The literature on clustering weighted or unweighted graphs (also often referred to as *community detection*, by analogy with social networks) is ample, and we refer the reader on this topic to surveys by Fortunato (2010) and Von Luxburg (2007), and to Section 3.7 on factorization-based methods.

3.6 Numerical Experiments

We present numerical experiments to highlight the benefits of our method. For efficiency reasons, we use the serial proximal descent algorithm (Algorithm 2).

3.6.1 Synthetic Data

Covariance matrix estimation. We draw N vectors $x_i \sim \mathcal{N}(0, \Sigma)$ for a block diagonal covariance matrix $\Sigma \in \mathbb{R}^{n \times n}$. We use r blocks of random sizes and of the form $vv^\top$ where the entries of v are drawn i.i.d. from the uniform distribution on $[-1, 1]$. Finally, we add Gaussian noise $\mathcal{N}(0, \sigma^2)$ on each entry. In our experiments $r = 5$, $N = 20$, $d = 100$, $\sigma = 0.6$. We apply our method (SPLR), as well as trace norm regularization (LR) and ℓ_1 norm regularization (SP) to the empirical covariance matrix, and report average results over ten runs. Figure 3.1 shows the RMSE normalized by the norm of Σ for different values of τ and γ. Note that the effect of the mixed penalty is visible as the minimum RMSE is reached inside the (τ, γ) region. We perform, on the same data, separate cross-validations on (τ, γ) for SPLR, on τ for LR and on γ for SP. We show in Figure 3.2 the supports recovered by each algorithm, the output matrix of LR being thresholded in absolute value. The support recovery demonstrates how our approach discovers the underlying patterns despite the noise and the small number of observations.

3.6.2 Real Data Sets

Protein Interactions. We use data from Hu et al. (2009), in which protein interactions in Escherichia coli bacteria are scored by strength in $[0, 2]$. The data is, by nature, sparse. In addition to this, it is often suggested that interactions between two proteins are governed by a small set of factors, such as surface accessible amino acid side chains Bock and Gough (2001), which motivates the estimation of a low rank representation. Representing the data as a weighted graph, we filter to retain only the 10% of all 4394 proteins that exhibit

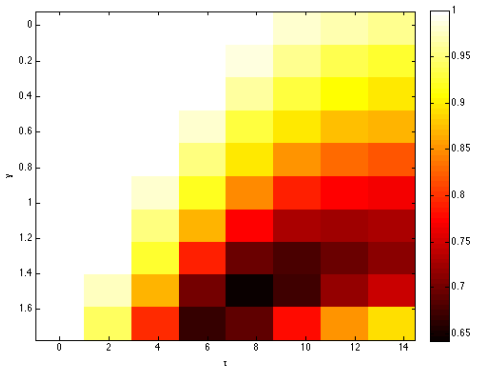

Figure 3.1: Covariance estimation. Cross-validation: normalized RMSE scores (SPLR)

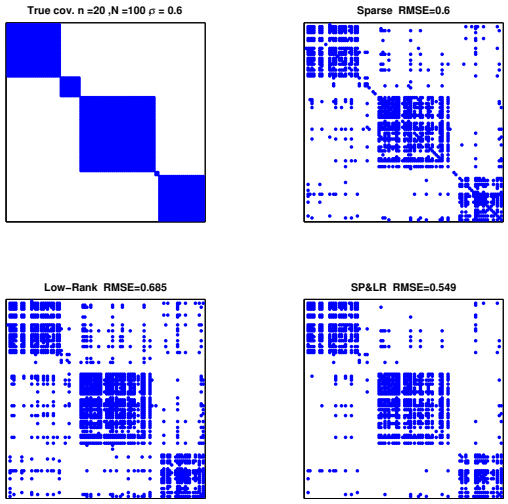

Figure 3.2: Covariance estimation. Support of Σ (top left), and of the estimates given by SP (top right), LR (bottom left), and SPLR (bottom right)

the most interactions as measured by weighted degree. We corrupt 10% of entries of the adjacency matrix selected uniformly at random by uniform noise in $[0, \eta]$. Parameters are selected by cross-validation and algorithms are evaluated using mean RMSE between estimated and original adjacency matrices over 25 runs. RMSE scores are shown in Table 3.1 and show the empirical superiority of our approach (SPLR).

η	SPLR	LR	SP
0.1	**0.0854** ±0.012	0.1487 ±0.02	0.1023 ±0.02
0.2	**0.2073** ± 0.03	0.2673 ± 0.3	0.2484 ± 0.03
0.3	**0.3105** ± 0.03	0.3728 ± 0.03	**0.3104** ± 0.02

Table 3.1: Prediction of interactions in Escherichia coli. Mean normalized RMSE and standard deviations.

Social Networks. We have performed experiments with the Facebook100 data set analyzed by Traud et al. (2012). The data set comprises all friendship relations between students affiliated to a specific university, for a selection of one hundred universities. We select a single university with 41554 users and filter as in the previous case to keep only the 10% users with highest degrees. In this case, entries are corrupted by impulse noise: a fixed fraction σ of randomly chosen edges are flipped, thus introducing noisy friendship relations and masking some existing relations. The task is to discover the noisy relations and recover masked relations. We compare our method to standard baselines in link prediction (Liben-Nowell and Kleinberg, 2007). Nearest Neighbors (NN) relies on the number of common friends between each pair of users, which is given by Y^2 when Y is the noisy graph adjacency matrix. Katz's coefficient connects a pair of nodes according to a score based on the number of paths connecting them, emphasizing short paths. Results are reported in Table 3.2 using the area under the ROC curve (AUC). SPLR outperforms LR but also NN and Katz which do not directly seek a low rank representation.

σ	SPLR	LR	NN	Katz
5 %	0.9293	0.9291	0.7680	**0.9298**
10 %	**0.9221**	0.9174	0.7620	0.9189
15 %	**0.9117**	0.9024	0.7555	0.9068
20 %	**0.8997**	0.8853	0.7482	0.8941

Table 3.2: Facebook denoising data. Mean AUC over 10 simulation runs. All standard deviations are lower than $3 \cdot 10^{-4}$.

3.7 Discussion

In this last section, we discuss various aspects of the proposed methods, as well as related work and perspectives.

3.7.1 Other Loss Functions

The methods presented in this chapter can be seamlessly extended to non-square matrices, which can arise, for instance, from adjacency matrices of bipartite graphs. Our work also applies to a wide range of other losses. A useful example that links our work to the matrix completion framework is when linear measurements of the target matrix or graph are available, or can be predicted as in (Richard et al., 2010). In this case, the loss can be defined in the feature space. Due to the low rank assumption, our method does not directly apply to the estimation of precision matrices often used for Gaussian graphical model structure learning (Friedman et al., 2008), and the applications of conditional independence structures generated by low rank and possibly sparse models is to be discussed. Note that the trace norm constraint is vacuous for some special classes of positive semi-definite matrices. For instance, it is not useful for estimating a correlation matrix as, in this case, the trace is always equal to the dimension.

3.7.2 Optimization

Other optimization techniques can be considered for future work. A trace norm constraint alone can be taken into account without projection or relaxation into a penalized form by casting the problem as a semi-definite program as proposed by Jaggi (2013). The special form of this semi-definite program can be leveraged to use the efficient resolution technique from (Hazan, 2008)., based on the Frank-Wolfe algorithm. This method applies to a differentiable objective whose curvature determines the performances. Extending these methods with projection onto the ℓ_1 ball or a sparsity-inducing penalty could lead to interesting developments.

3.7.3 Geometry

A major drawback of the proposed method is that the unit ball of the proposed penalty has extreme points that are either sparse or low rank, but that are not usually both. In a recent analysis (Drusvyatskiy et al., 2014) focused on the geometry of the unit ball of the combined penalty, a more precise characterization of the extremal points is given: any extremal point W has to satisfy

$$\frac{r(r+1)}{2} \leq k+1,$$

where r and k are the rank and the ℓ_0-norm of W, respectively. In addition, they show that the unit ball of the combined penalty does not have more vertices than the ℓ_1-norm unit ball. This is not surprising: as we have mentioned in Section 2.4.3, the unit ball for the trace norm has no vertices, although it has higher-order non-smooth points. However, it is not clear how relevant the number of vertices (i.e., of maximally non smooth points) is for recovery. A further analysis of higher-order points of the unit ball of the combined norm could be fruitful.

We have provided an analysis for recovery of fully-observed but noisy matrices. An interesting alternate problem consists in characterizing the sample complexity, both with or without noise, as shown in Chapter 2. As was shown recently based on Gaussian width arguments (Oymak et al., 2012), the sample complexity is not reduced when combining ℓ_1-norm and trace norm for exact or robust recovery with Gaussian designs. This suggests that the experimental design matters significantly when designing regularizers.

Another interesting question consists in determining tighter convex relaxations to the rank and the sparsity. In particular, one may ask what is the convex hull of

$$\left\{W \in \mathbb{R}^{d_1 \times d_2} : \operatorname{rank}(W) \leq r, \|W\|_0 \leq k\right\}$$

intersected with the ℓ_∞-norm ball, or over other balls. However, it may also be that there are some more interesting joint measures of rank and sparsity to relax than the intersection of these two classical manifolds. The *ranksity* index (Richard et al., 2013) provides such an alternative measure, along with a convex relaxation based on the trace norm. This convex penalty has the benefit of being non-differentiable at points which are both low rank and sparse. Although such a modeling work can also be done through atomic norms in principle (e.g., considering rank one atoms with limited sparsity), it remains unclear how the resulting norms can be used in practice.

3.7.4 Factorization Methods

A related task is finding low rank factorizations of matrices of the form UV^T (Srebro, 2004; Srebro et al., 2005), thus jointly optimizing in $U, V \in \mathbb{R}^{d\times r}$ loss functions of the form $\ell((U,V),Y) = \|UV^T - Y\|_F^2$ for some target maximum rank r. This implicitly encodes the low rank constraint which leads to efficient optimization schemes, and allows for interpretability as estimated (U,V) pairs can be considered as latent factors. Although optimization with the trace norm or the max norm implicitly assumes such a factorization, factorization-based method work with the factorization directly. Although formulations are usually convex in U or V, they are not in general jointly convex and optimization procedures can get stuck in local minima. However, these methods have been associated with good empirical performances (Xu et al., 2003; Koren, 2008). Factorization methods parameterized in U and V are not very practical to control the sparsity of $W = UV^T$. However, one may alternatively seek to control the sparsity of the factors U and V, which can lead to sensible models. In the following, we give more background on these methods, which could potentially be of interest to define regularizers based on factorizations.

Sparse Factors

Nonnegative Matrix Factorization (NMF) (Lee et al., 1999) imposes nonnegativity constraints on the coefficients of U and V to enhance interpretability by allowing only for additive effects and tends to produce sparse factor matrices U, V, although this a rather indirect effect. There is no strong guarantee on the sparsity achieved by NMF nor is it easy to set the target sparsity. Different methods for sparse NMF have been proposed (Hoyer, 2004a; Kim and Park, 2008), where sparsity inducing regularizers are applied on the coefficients of the factors. Sparse matrix factorizations have also been proposed without the positivity constraint, such as for sparse coding (Hoyer, 2002), or for extending the classical PCA and finding sparse directions that maximize the variance of the projection. This last problem is often referred to as *sparse principal component analysis*. SPCA (Zou et al., 2004) penalizes the ℓ_1 norm of the principal components and can be reduced to solving independent elastic-nets. A different formulation using SDP programming is introduced by d'Aspremont et al. (2007) with good empirical results. In addition, some methods give up up orthogonality between the components (Mackey, 2009).

An important special case of matrix factorizations is that of rank one factorizations. The detection of *cliques* within adjacency matrices correspond to factorizations of the form $\mathbf{1}_S\mathbf{1}_S^T$ where $S \subset [d]$ is a set of coordinates. A penalty similar to the one we proposed has been used in (Ames and Vavasis, 2011) to detect cliques, although with a different loss, which can be shown to lead to exact recovery of the support of the clique in some circumstances. These types of models are discussed more in the second part of this thesis, in Chapter 5.

Interpolation of Nuclear Norms

The combined penalty $\gamma\|.\|_1 + (1-\gamma)\|.\|_*$ that we propose interpolates between two nuclear norms: the trace norm, and the ℓ_1-norm. Indeed, from duality with operator norms,

the ℓ_1-norm can be expressed in nuclear form as well:

$$\|W\|_1 = \|W\|_{1\to\infty} = \inf\left\{\sum_i |\sigma_i| \,:\, W = \sum_i \sigma_i b_i a_i^T,\ \|b_i\|_1 = \|a_i\|_1 = 1\right\}.$$

The linear combination that we propose is surely not the only way to interpolate between the ℓ_1-norm and the trace norm, and this suggests the following question: what are interesting alternative interpolation methods? A potential choice consists in using semidefinite programming relaxations: consider

$$F(X) = \gamma\|X\|_1 + (1-\gamma)\operatorname{Tr}(X),$$

such that for any $u \in \mathbb{R}^n$,

$$F(uu^T) = \gamma\|u\|_1^2 + (1-\gamma)\|u\|_2^2.$$

This allows to define the following norms.

Definition 10.

$$\|X\|_{uv} = \frac{1}{2}\min_{UV^T=X}\sum_{i\geq 1}\left\{F(u_i u_i^T) + F(v_i v_i^T)\right\},$$

$$\|X\|_{SDP} = \frac{1}{2}\min_{[Z_1 X; X^T Z_2]\geq 0}\left\{F(Z_1) + F(Z_2)\right\}.$$

As usual, the dimensions of U and V are not restricted. Although $\|\cdot\|_{\text{uv}}$ interpolates between the trace norm and the ℓ_1-norm when $\gamma \in [0,1]$, this norm cannot usually be computed. However, both the SDP norm and the sum of ℓ_1-norm and trace norm are lower bounds on $\|\cdot\|_{\text{uv}}$, and are thus convex relaxation which can be used in practice. As it turns out, $\|\cdot\|_{\text{SDP}}$ does not interpolate correctly at $\gamma = 1$: although for $\gamma = 0$, $\|X\|_{\text{SDP}} = \|X\|_{\text{uv}} = \|X\|_*$, there exists matrices W such that $\|W\|_{\text{SDP}} < \|W\|_{\text{uv}} = \|W\|_1$ for $\gamma = 1$ (consider, for instance, $W = [1\,1; 1\,-1] \in \mathbb{R}^{2\times 2}$, which has $\|W\|_1 = 4$, but $\|W\|_{\text{SDP}} < 3 < \|W\|_1$.) As a consequence, this does not provide an acceptable interpolation. However, we believe that investigating other interpolations and relaxations along these lines could be of interest.

Factorization Constants

Straying away from the idea of interpolation between rank and sparsity, one may seek to define a penalty directly, using factorizations. Like the trace norm and the max norm, many more norms can be defined through factorizations.

Proposition 15. *Consider normed vector spaces* $(\mathbb{R}^{d_2}, \|\cdot\|_X)$, $(Z, \|\cdot\|_Z)$ *and* $(\mathbb{R}^{d_1}, \|\cdot\|_Y)$. *Then,*

$$\gamma_Z(W) = \inf\{\|V\|_{X\to Z}\|U\|_{Z\to Y} \,:\, W = UV\}$$

is a norm over $\mathbb{R}^{d_1\times d_2}$.

Again, the number of columns of U and V in the factorizations is left unbounded. This type of norm is referred to as a *factorization constant*, as it measures how well a linear operator W can be factorized through some latent space Z. This suggests the following type of norm.

Proposition 16. *Consider for $W \in \mathbb{R}^{d\times d}$ the functionals*

$$q(W) = \min_{UV^T=W} \|U\|_1\|V\|_1, \quad c_r(W) = \min_{\substack{U\in\mathbb{R}^{d\times r}\\ V\in\mathbb{R}^{d\times r}\\ UV^T=W}} \|U\|_1\|V\|_1.$$

Then, q is a norm over $\mathbb{R}^{d\times d}$, but c_2 is not a norm over $\mathbb{R}^{2\times 2}$.

The proof is deferred to Section 3.8.2. Although leaving the number of factors unbounded yields a norm which can measure sparsity of factorizations, this norm cannot be computed in practice. On this other hand, capping the number of factors even at the ambient dimension does not necessarily leads to a convex function. Again, we believe that variations on this type of approach could be developed and yield interesting results.

3.8 Proofs

3.8.1 Sketch of Proof of Proposition 13

Proof. For any W in $\mathcal{W}$ and by optimality of $\widehat{W}$,

$$-2\langle \widehat{W} - W, W^\star \rangle \leq 2\alpha \|\widehat{W} - W\|_* \|\epsilon\|_{\text{op}} + 2(1-\alpha)\|\widehat{W} - W\|_1 \|\epsilon\|_\infty + \tau(\|W\|_* - \|\widehat{W}\|_*) + \gamma(\|W\|_1 - \|\widehat{W}\|_1) + \|W\|_F^2 - \|\widehat{W}\|_F^2$$

for any $\alpha \in [0;1]$. The assumptions on τ, γ and triangular inequality lead to the first bound.

Let $r = \text{rank}(W)$, $k = \|W\|_0$, $W = \sum_{j=1}^r \sigma_j u_j v_j^\top$ the SVD of W, $W = \Theta \circ |W|$, where $\Theta = \text{sgn}(W)$, and $\Theta^\perp \in \{0,1\}^{d \times d}$ the complementary sparsity pattern. We use $P_{W_1^\perp}$ (resp. $P_{W_2^\perp}$) to denote the projection operator onto the orthogonal of the left (resp. right) singular space of W. We also note $\mathcal{P}_W(X) = X - P_{W_1^\perp} X P_{W_2^\perp}$ such that $X = \mathcal{P}_W(X) + P_{W_1^\perp} X_{W_2^\perp}$.

Any element V of the subgradient of the convex function $W \mapsto \tau\|W\|_* + \gamma\|W\|_1$ can be decomposed as

$$V = \tau\left(\sum_{j=1}^r u_j v_j^\top + P_{W_1 \perp} Q_* P_{W_2 \perp}\right) + \gamma\left(\Theta + Q_1 \circ \Theta^\perp\right)$$

for Q_1, Q_* with $\|Q_*\|_{\text{op}} \leq 1$, $\|Q_1\|_\infty \leq 1$, which can be chosen such that

$$\left\langle V, \widehat{W} - W \right\rangle = \tau\left\langle \sum_{j=1}^r u_j v_j^\top, \widehat{W} - W \right\rangle + \tau\|P_{W_1 \perp} \widehat{W} P_{W_2 \perp}\|_* + \gamma\langle \Theta, \widehat{W} - W \rangle + \gamma\|\Theta^\perp \circ \widehat{W}\|_1 .$$

By monotonicity of the subdifferential and optimality conditions,

$$2\langle \widehat{W} - W^\star, \widehat{W} - W \rangle \leq 2\langle \epsilon, \widehat{W} - W \rangle - \tau\left\langle \sum_{j=1}^r u_j v_j^\top, \widehat{W} - W \right\rangle - \tau\|P_{W_1^\perp} \widehat{W} P_{W_2^\perp}\|_* - \gamma\langle \Theta, \widehat{W} - W \rangle - \gamma\|\Theta^\perp \circ \widehat{W}\|_1 .$$

Decompose

$$\epsilon = \alpha\left(\mathcal{P}_W(\epsilon) + P_{W_1^\perp} \epsilon P_{W_2^\perp}\right) + (1-\alpha)\left(|\Theta| \circ \epsilon + \Theta^\perp \circ \epsilon\right).$$

Using results on dual norms, we have

$$|\langle M_1, M_2 \rangle| \leq \|M_1\|_* \|M_2\|_{\text{op}}$$
$$|\langle M_1, M_2 \rangle| \leq \|M_1\|_1 \|M_2\|_\infty$$

for all $M_1, M_2 \in \mathbb{R}^{n \times n}$ and hence,

$$\begin{aligned}\langle \epsilon, \widehat{W} - W \rangle \leq{}& \alpha\|\mathcal{P}_W(\epsilon)\|_F \|P_{W_1}(\widehat{W} - W) P_{W_2}\|_F \\ &+ \alpha\|P_{W_1^\perp} \epsilon P_{W_2^\perp}\|_{\text{op}} \|P_{W_1^\perp} \widehat{S} P_{W_2^\perp}\|_* \\ &+ (1-\alpha)\|\Theta \circ \epsilon\|_F \|\Theta \circ \left(\widehat{W} - W\right)\|_F \\ &+ (1-\alpha)\|\Theta^\perp \circ \epsilon\|_\infty \|\Theta^\perp \circ \widehat{W}\|_1 .\end{aligned}$$

Using

$$\|\mathcal{P}_W(\epsilon)\|_F \leq \sqrt{2\,r}\|\epsilon\|_{op}, \quad \|\Theta \circ \epsilon\|_F \leq \sqrt{k}\|\epsilon\|_\infty$$

leads for $\tau \geq 2\alpha\,\|\epsilon\|_{\text{op}}$ and $\gamma \geq 2(1-\alpha)\,\|\epsilon\|_\infty$ to

$$\begin{aligned}&\|\widehat{W} - W^\star\|_F^2 + \|\widehat{W} - S\|_F^2 \\ &\qquad \leq \|W - W^\star\|_F^2 + \left(\tau\sqrt{r}(\sqrt{2}+1) + 2\gamma\sqrt{k}\right)\|\widehat{W} - W\|_F.\end{aligned}$$

Using $\beta x - x^2 \leq \left(\frac{\beta}{2}\right)^2$, we obtain

$$\|\widehat{W} - W^\star\|_F^2 \leq \|W - W^\star\|_F^2 + \frac{1}{4}\left(\sqrt{r}\tau(\sqrt{2}+1) + 2\sqrt{k}\gamma\right)^2$$

and setting $W = W^\star$ gives the result. □

3.8.2 Proof of Proposition 16

Proof. q is a norm from the previous proposition, as the ℓ_1-norm is an operator norm. Let $X = uv^T \in \mathbb{R}^{d\times d}$ a rank one matrix, then for any $r \geq 1$,

$$c_r(X) = \|u\|_1\|v\|_1.$$

Let $d = 2$, $E_{1,1} = \text{diag}(1,0), E_{2,2} = \text{diag}(0,1) \in \mathbb{R}^{2\times 2}$, and I_2 be the identity matrix of size 2. We will begin by showing that $c_2(I_2) = 4$. Then, this will imply $c_2(E_{1,1}) + c_2(E_{2,2}) < c_2(E_{1,1}+E_{2,2})$ and as a consequence, c_2 does not satisfy the triangle inequality. For any decomposition such that $UV^T = I_2$, it must hold that $V^T = U^{-1}$. Since $\|X\|_1 = \|X^T\|_1$, we have

$$c_2(I_2) = \min_{\substack{U\in\mathbb{R}^{2\times 2}\\ \det(U)\neq 0}} \|U\|_1\|U^{-1}\|_1.$$

Consider

$$d_2(I_2) = \min_{\substack{U=\text{diag}(d_1,d_2)\\ \det(U)\neq 0}} \|U\|_1\|U^{-1}\|_1.$$

Assume without loss of generality that $d_i > 0$ for $i \in \{1,2\}$. We have

$$\|U\|_1\|U^{-1}\|_1 = (d_1+d_2)\,(1/d_1 + 1/d_2) = 2 + \frac{d_1}{d_2} + \frac{d_2}{d_1}.$$

Since $(d_1 - d_2)^2 \geq 0$, we have $\frac{d_1}{d_2} + \frac{d_2}{d_1} \geq 2$ and hence, $\|U\|_1\|U^{-1}\|_1 \geq 4$. This implies $d_2(I_2) = 4$, achived at $U = I_2$. Now, let

$$U = \begin{bmatrix} a & c \\ d & b \end{bmatrix}.$$

We have $\|U\|_1\|U^{-1}\|_1 = \|U\|_1^2/|\det(U)|$, which is minimum either when U is diagonal ($c = d = 0$) or anti-diagonal ($a = b = 0$). Without of loss generality, we can exclude this last case as $\|.\|_1$ is invariant to permutations. This implies that $d_2(I_2) = c_2(I_2)$, and thus $c_2(I_2) = 4$. □

4

Convex Localized Multiple Kernel Learning

Contents

4.1 Introduction . . . 82
4.1.1 Linear MKL . . . 82
4.1.2 Hinge Loss Aggregation . . . 83
4.1.3 Related Work . . . 84
4.1.4 Outline . . . 86
4.1.5 Notation . . . 86
4.2 Generalized Hinge Losses . . . 86
4.2.1 Representer Theorem . . . 88
4.2.2 Universal Consistency . . . 89
4.3 ℓ_p-norm Aggregation of Hinge Losses . . . 91
4.3.1 Dual Problem . . . 91
4.3.2 Regularization . . . 93
4.3.3 Link Functions . . . 94
4.4 Experiments . . . 94
4.4.1 UCI Datasets . . . 95
4.4.2 Image Classification . . . 96
4.5 Discussion . . . 97
4.6 Proofs . . . 98
4.6.1 Proof of Lemma 2 . . . 98
4.6.2 Proof of Lemma 3 . . . 98
4.6.3 Proof of Lemma 4 . . . 99
4.6.4 Proof of Lemma 5 . . . 99
4.6.5 Proof of Theorem 3 . . . 100

4.1 Introduction

Kernel-based methods have become a classical tool of the trade in machine learning. In addition to good empirical performance in a wide range of situations, there is a strong theoretical background behind some kernel machines such as SVMs (Steinwart and Christmann, 2008), as well as mature algorithms (Platt, 1999) and implementations (Chang and Lin, 2011). The kernel is traditionally either selected from generic parametric off-the-shelf kernels (e.g., Gaussian or polynomial) using some sort of cross-validation, or hand-crafted by domain specific experts through an expensive empirical process. Recently, metric learning has become an increasingly active research field (Kulis, 2013), with the goal of learning a kernel suitable for a given task. Multiple kernel learning (MKL) has appeared as an alternative, in between manual kernel selection and metric learning, where a weighting of different representations of the data points is learnt at the same time as a large margin classifier (Lanckriet et al., 2002; Bach et al., 2004; Gönen and Alpaydın, 2011). Representations may correspond to multiple scales, or to more orthogonal types of information (e.g., in images, different types of geometric or color information).

Although this has been successful, the linear kernel mixture in MKL is *global* in the sense that the weighting of the kernels is independent of training data points. Recently, alternatives have been proposed where the kernel mixture is *local* Gönen and Alpaydın (2013), in the sense that the weighting of each kernel depends on where it is evaluated. These approaches are referred to as either *localized*, or *data-dependent* multiple kernel learning. Indeed, in many problems, noise levels and discrimination ability in different views may vary across the input feature space, and this observation is a strong incentive for further exploration of localization with MKL methods. This has been approached in different ways. Parametric forms of data-dependent kernel weighting have been proposed by Gönen and Alpaydin (2008). Nonparametric data-dependent kernel weighting methods have also been proposed, including large-margin approaches (Yang et al., 2010), other types of losses and penalizations (Bi et al., 2004; Cao et al., 2009), and ensemble methods (Gehler and Nowozin, 2009).

These methods are either non-convex, lack theoretical guarantees, or do not correspond to large margin approaches. In this chapter, we propose a family of large-margin methods that are both convex and theoretically grounded for combining kernels in a data-dependent manner.

4.1.1 Linear MKL

Consider the setting of binary supervised classification in $\mathbb{R}^d$, with N training data points $(x_1, y_1), \ldots, (x_N, y_N)$ independently drawn from a distribution P over $\mathbb{R}^d \times \{-1, 1\}$. We consider a family $(K_m)_{1 \leq m \leq M}$ of kernels, with associated feature mappings $\phi_m : \mathbb{R}^d \to H_m$ for $1 \leq m \leq M$, where H_m is the corresponding RKHS. For kernel mixing weights $s = (s_1, \ldots, s_M) \in \mathbb{R}_+^M$, the global mixture kernel $K = \sum_{m=1}^M s_m K_m$ corresponds to the feature mapping

$$\Phi(x) = \left[\sqrt{s_1}\,\phi_1(x), \ldots, \sqrt{s_M}\,\phi_M(x)\right]^T,$$

and is associated with the product RKHS $H = H_1 \times \ldots \times H_M$. Plugging K into the standard SVM optimization problem yields

$$\min_{\substack{\omega=(\omega_1,\ldots,\omega_M)\in H,\\ s\geq 0}} \left[\frac{1}{2} \sum_{m=1}^{M} \|\omega_m\|_2^2 + C \sum_{i=1}^{N} \left(1 - y_i \sum_{m=1}^{M} \sqrt{s_m} \langle \omega_m, \phi_m(x_i) \rangle \right)_+ \right], \tag{4.1}$$

where $s \geq 0$ denotes a componentwise nonnegativity constraint (which ensures that the mixture is a valid kernel), and $(\cdot)_+$ denotes the positive part. Although this objective is non-convex, the change of variable $\sqrt{s_m}\omega_m = w_m$ leads to an objective that is jointly convex in (w, s). In addition, a regularizer $\Omega : \mathbb{R}^M \to \mathbb{R}_+ \cup \{\infty\}$ can be used to control the mixture coefficients - classical choices include ℓ_p-norms for $p \geq 1$, indicator functions of the corresponding unit balls, or Bregman divergences. In particular, an ℓ_1-norm constraint can force the weights associated to the kernels to sum to unity. The resulting problem is then

$$\min_{\substack{w=(w_1,\ldots,w_M)\in H,\\ s\geq 0}} \left[\frac{1}{2} \sum_{m=1}^{M} \frac{\|w_m\|_2^2}{s_m} + C \sum_{i=1}^{N} \left(1 - y_i \sum_{m=1}^{M} \langle w_m, \phi_m(x_i) \rangle \right)_+ + \Omega(s) \right].$$

The partial convex dual to this problem (with respect to w only) is of the form

$$\min_{s\geq 0} \max_{\alpha\in[0,C]^N} \left[\mathbf{1}^T \alpha - \frac{1}{2} \alpha^T \left(\sum_{m=1}^{M} s_m H_m \right) \alpha + \Omega(s) \right]$$

with $H_m = \operatorname{diag}(y) K_m \operatorname{diag}(y)$, where $\operatorname{diag}(y)$ is the diagonal matrix with diagonal elements $y \in \mathbb{R}^N$, and $\mathbf{1}$ is the all ones vector. This saddle point problem has been solved using various methods, and many optimization methods have been developed in order to address certain choices of regularizers on s. This variant of MKL is usually referred to as *linear MKL*, and results in decision functions of the form

$$f(x) = \sum_{i=1}^{N} y_i \alpha_i \left[\sum_{m=1}^{M} s_m K_m(x_i, x) \right],$$

which can be augmented by a constant bias term. This form of decision function is not localized, as the relative weights of the kernels are the same at all the support vectors, meaning that the *same* kernel combination is used everywhere in the feature space.

4.1.2 Hinge Loss Aggregation

In this work, we set out to learn decision functions of the more general form

$$f(x) = \sum_{i=1}^{N} \sum_{m=1}^{M} y_i \alpha_{i,m} K_m(x_i, x) \tag{4.2}$$

with $\alpha \in \mathbb{R}^{N\times M}$. While the solution to the linear MKL problem can be written in this form with $\operatorname{rank}(\alpha) = 1$, we are interested in higher rank solutions which exhibit localization, in the sense that the relative weights of the kernels vary depending on the support vectors. Informally, this corresponds to a kernel combination that is different in different regions

of the feature space. In order to induce such a structure on the solution while maintaining the large-margin interpretation, we consider optimization problems of the form

$$\min_{\omega=(\omega_1,\ldots,\omega_M)\in H}\left[\frac{1}{2}\sum_{m=1}^{M}\|\omega_m\|_2^2 + C\sum_{i=1}^{N}\ell\left(y_i,\langle\omega_1,\phi_1(x_i)\rangle,\ldots,\langle\omega_M,\phi_M(x_i)\rangle\right)\right], \tag{4.3}$$

where $\ell(y,t_1,\ldots,t_M)$ is a generalization of the hinge loss to a collection of classifiers $(t_1,\ldots,t_M)$ with $t_i = \langle\omega_i,\phi_i(x)\rangle$ for a data point x. For certain classes of losses ℓ, problems of the form (4.3) can be shown to lead to universally consistent estimators. Among these estimators, some choices of ℓ may result in a localized solution of the form (4.2), and we study examples of such choices in this chapter.

More precisely, with $z = (z_1,\ldots,z_M)$ the vector of hinge losses such that $z_m = (1 - yt_m)_+$, we propose to use $\ell(y,t_1,\ldots,t_M) = \|z\|_p$, the ℓ_p-norm of z. With q such that $1/p + 1/q = 1$, this leads to a dual problem of the form

$$\max_{\alpha\geq 0}\left[\mathbf{1}^T\alpha\mathbf{1} - \frac{1}{2}\sum_{m=1}^{M}\alpha_{\cdot,m}^T H_m \alpha_{\cdot,m}\right] \text{ s.t. } \|\alpha_{i,\cdot}\|_q \leq C \text{ for } i\in[N],$$

where for any matrix $\alpha\in\mathbb{R}^{N\times M}$, we denote by $\alpha_{i,\cdot}$ the i-th row of X, and by $\alpha_{\cdot,m}$ the m-th column of X. Hence, this choice of loss results in a coupling of the SVMs associated to the different kernels to be combined, which can be adjusted through the choice of p.

In case where $p = 1$, the problem consists in training M independent SVMs and averaging their decision functions, while a coupling constraint between the weights associated to different kernels is present in the case $p = \infty$. Intermediate values of p correspond to intermediate amounts of couplings between the kernels. The following question is at the core of this work: what is the *best amount of coupling* between the kernels? From a theoretical perspective, we show that all couplings lead to universally consistent classifiers. From a practical perspective, we show on various experimental benchmarks that low or no coupling between the kernels seems to be best, and allow to outperform existing methods for localized MKL.

4.1.3 Related Work

A localized MKL (LMKL) was proposed by Gönen and Alpaydin (2008), with localized weighting of the kernels according to the data-dependent parametric form

$$f(x) = \sum_{i=1}^{N} y_i\alpha_i\left[\sum_{m=1}^{M}\eta_m(x)\,K_m(x_i,x)\right] \text{ with } \eta_m(x) = \frac{\mathbf{e}^{a_m+\langle b_{\cdot,m},x\rangle}}{\sum_{n=1}^{M}\mathbf{e}^{a_n+\langle b_{\cdot,n},x\rangle}},$$

and $a\in\mathbb{R}^M, b\in\mathbb{R}^{d\times M}$ to be learned from the data in addition to α. The *gating functions* (η_m) allow to activate different kernels depending on the input point. Although a general kernel can in principle be used instead of an explicit feature representation to define the gating functions, the original training procedure is based on gradient descent with respect to the gating parameters, which precludes the use of kernelized gating functions. This form of decision function is different from what we consider in this work: in LMKL, although the weighting of the individual kernels depends on the input point, the support vectors are identical for all the kernels. In particular, at a fixed input point, this corresponds as in

MKL to a rank one weighting. Indeed, when the original vector representation used for gating has dimension d, LMKL estimates a model of dimension $N+M(d+1)$, which can be much less than NM. In addition, although the decision functions of LMKL can be obtained by iteratively resorting to a black-box SVM solver, the problem of jointly estimating the gating weights and SVM weights is non-convex, and the resulting estimator may result from high variance due to the effect of initialization.

The CG-Boost method (Bi et al., 2004) consists in plugging a classifier of the form (4.2) into the hinge loss, together with ℓ_2-norm regularization, leading to

$$\min_{\alpha \in \mathbb{R}_+^{N \times M}} \left[\frac{1}{2} \sum_{m=1}^{M} \|\alpha_{\cdot,m}\|_2^2 + C \sum_{i=1}^{N} \left(1 - y_i \sum_{m=1}^{M} \sum_{j=1}^{N} \alpha_{j,m} K_m(x_i, x_j) \right)_+ \right].$$

This can be interpreted as a linear SVM with NM features that correspond to kernel evaluations with all elements of the training set. In particular, the regularization is independent of the original geometry of the kernels, and there is no direct large margin interpretation.

The per-sample MKL (PS-MKL) (Yang et al., 2010, 2009) is another localized ("per-sample") MKL approach with decision functions of the form

$$f(x) = \sum_{i=1}^{N} y_i \sum_{m=1}^{M} \alpha_i \beta_m(x_i, x) K_m(x_i, x)$$

for $\alpha \in \mathbb{R}_+^N$. In order to obtain a tractable problem, the weighting is assumed for learning to be of the form $\beta_m(x_i, x_j) = \frac{1}{2}\left(\beta_{i,m} + \beta_{j,m}\right)$, while for testing, the authors consider either $\beta_m(x_i, x) = \beta_{i,m}/M$ or $\beta_m(x_i, x) = \frac{1}{2}\left(\beta_{i,m} + 1/M\right)$, which is to say that the weighting is uniformly affected by unseen examples. With $\beta \in \mathbb{R}_+^{N \times M}$, this form of the decision function is plugged into the classical SVM formulation, as when deriving the classical linear MKL formulation in (4.1). Unlike CG-Boost, this leads to a regularization that is dependent on the kernels. However, the change of variable trick used with MKL to transform the problem into a convex optimization problem does not apply here, and this leads to a saddle-point problem, to be solved using alternating optimization over β and α. This corresponds to the same class of decision functions that we consider, albeit with a different parameterization which does not lead to a convex formulation.

Outside of the realm of SVMs, Cao et al. (2009) have introduced the Heterogeneous Feature Machine (HFM), where localized decision functions of the form (4.2) are trained using a group-lasso regularized logistic loss with groups correspond to training samples, resulting in a problem of the form

$$\min_{\alpha \in \mathbb{R}^{N \times M}} \sum_{i=1}^{N} \log\left(1 + e^{-y_i f(x_i)}\right) + \lambda \sum_{i=1}^{N} \|\alpha_{i,\cdot}\|_2, \tag{4.4}$$

with f as in (4.2). Although this is also similar to our approach in terms of the class of decision functions, this is not a large margin approach. In addition, as for all the previous methods, no theoretical guarantees are provided.

MKL is also related to multi-view learning (Blum and Mitchell, 1998), which is traditionally interested in problems where features can be divided into subsets (views), each of which is sufficient to learn a good predictor (such that there is redundancy between the

views). On the opposite, MKL can also be seen as an ensemble method, which are usually concerned with aggregating estimators, which may be potentially loosely or not correlated. This dual view is well reflected in the two methods that we propose. For $p = \infty$, we penalize the models associated to each kernel according to by how much they disagree on the data points, as is common in multi-view learning. On the other hand, for $p = 1$, the solution consists in averaging decision functions of independent SVMs, which can be seen as an ensemble method with base classifiers corresponding to SVMs with different kernels. Ensemble methods of similar sorts have been previously proposed, such as LP-β (Gehler and Nowozin, 2009) which consists in using independently trained SVMs as base learners, and finding a weighting of these through a LPboost-like penalized linear program (Demiriz et al., 2002). This has been shown to yield good results, but requires to tune both penalization in the base learners, and when combining these.

Finally, the problem of localized MKL has close ties to metric learning, where one seeks to learn a full $N \times N$ kernel matrix over the training set, and, possibly, to extrapolate to new data points (see, for instance, (Yang and Jin, 2006; Ong et al., 2005) and references therein). MKL appears as a special case of metric learning, over a parametric class of kernels, while localized MKL corresponds to a larger class of kernels. Using a localized MKL, we only need to learn NM parameters, as opposed to N^2 for learning a full kernel. Although there has been some interest in using MKL with a large number of kernels (Gehler and Nowozin, 2008), the number M of kernels remains small for a wide variety of applications in which metric learning is thus a more difficult task.

4.1.4 Outline

In Section 4.2, we study some properties of optimization problems of the form (4.3) for different families of losses. We give examples, along with a representer theorem. In addition, we give conditions under which these losses lead to universally consistent classifiers. In Section 4.3, we look into more detail at the ℓ_p-norm loss aggregation, and show that this leads to a localized MKL solution which can be computed using a convex program. Experimental results are presented in Section 4.4, and a discussion is provided in Section 4.5. This chapter is joint work with Antoine Poliakov.

4.1.5 Notation

Throughout the chapter, we denote $\{1,\dots,N\}$ by $[N]$, and, similarly, $\{1,\dots,M\}$ by $[M]$. For a vector $x \in \mathbb{R}^d$ for some d, we usually denote by $(x_1,\dots,x_d)$ the corresponding components. The classical hinge loss is $\ell_{\text{hinge}}(y,x) = (1-yx)_+$. The all ones vector is denoted by $\mathbf{1}$, while $\mathbf{1}(A)$ designates the indicator vector that is one when condition A is true, and zero otherwise.

4.2 Generalized Hinge Losses

In this section, our objective is to study some properties of the minimizer of the following optimization problem where we aggregate hinge losses associated to all M kernels:

$$\min_{\omega=(\omega_1,\dots,\omega_M)\in H} \left[\frac{1}{2} \sum_{m=1}^{M} \|\omega_m\|_2^2 + C \sum_{i=1}^{N} \ell\left(y_i, \langle \omega_1, \phi_1(x_i)\rangle, \dots, \langle \omega_M, \phi_M(x_i)\rangle\right) \right]. \tag{4.5}$$

Although we use the *aggregation* terminology, this differs from approaches which consider aggregation of estimators (Freund et al., 1999; Dalalyan and Tsybakov, 2007), while (4.5) consists in aggregating loss functions. We impose the following conditions on the loss function.

Definition 11. *A loss function $\ell : \{-1,1\} \times \mathbb{R}^M \to \mathbb{R}_+$ is a generalized hinge loss (GHL) if, for $y \in \{-1,1\}$,*

- *$\ell(y,\cdot)$ is Lipschitz with constant L_ℓ with respect to the sup norm $\|t\|_\infty = \max_{i\in[M]} |t_i|$: for any $t, t' \in \mathbb{R}^M$,*
$$|\ell(y,t) - \ell(y,t')| \leq L_\ell \|t - t'\|_\infty.$$
- *$\ell(y, t\mathbf{1}) = (1-yt)_+$, and there exists $\theta : \mathbb{R}^M \to \mathbb{R}$ such that for any $t \in \mathbb{R}^d$,*
$$\ell(y, t_1, \ldots, t_M) \geq (1 - y\theta(t_1, \ldots, t_M))_+.$$

Following classical terminology in generalized linear models, we refer to θ as the *link function*, and the probabilistic classifier associated to the optimization problem is defined as
$$f_\ell(x) = \theta\left(\langle \omega_1, \phi_1(x)\rangle, \ldots, \langle \omega_M, \phi_M(x)\rangle\right).$$
where $\omega = (\omega_1, \ldots, \omega_M) \in H$ is the minimizer of (4.5). Note that this is slightly more general than (4.2), which corresponds to a link function that is simply the sum. This allows for more flexibility, and still can be considered a localized classifier. In practice, we shall further require that $\ell(y,\cdot)$ be convex for any y. In this chapter, we consider two classes of GHLs, that are built either by combining hinge losses with scalar inputs, or feeding a combination of scalar inputs into the hinge loss.

Definition 12. *A GHL ℓ is of outer combination type if there exists $\mathcal{A} : \mathbb{R}^M \to \mathbb{R}_+$ such that*
$$\ell(y, t_1, \ldots, t_M) = \mathcal{A}((1-yt_1)_+, \ldots, (1-yt_M)_+).$$
Similarly, a GHL ℓ is of inner combination type if there exists $\mathcal{A} : \mathbb{R}^M \to \mathbb{R}$ such that
$$\ell(y, t_1, \ldots, t_M) = (1 - y\mathcal{A}(t_1, \ldots, t_M))_+.$$

In both cases, we refer to $\mathcal{A}$ as the corresponding *aggregation function*. Note that the aggregation function must satisfy the scaling condition $\mathcal{A}(t, \ldots, t) = t$ for $t \in \mathbb{R}$. In particular, GHLs of outer combination type can be interpreted as generalizations of the classical arithmetic average. By definition, any GHL ℓ is lower bounded by a GHL of inner combination type, and the corresponding aggregation function is a link function for ℓ. Although it may appear from this fact that inner combinations should always be preferred, general GHLs or GHLs of outer combination type may lead to more tractable optimization problems, as is already the case with the hinge loss relaxation to the classification loss. When $\mathcal{A}$ is linear, we recover the two extreme cases that are averaging the decisions of individual SVMs, and training a SVM on the average kernel.

Lemma 1. *Let ℓ be of outer combination type with $\mathcal{A} = [a_1, \ldots, a_M]$ linear, then the solution of (4.5) is $w = (a_1\bar{w}_1, \ldots, a_M\bar{w}_M)$ where $\bar{w}_m$ is the solution of the SVM trained with kernel K_m only. Similarly, if ℓ is of inner combination type with the same $\mathcal{A}$, then the solution of (4.5) is the solution of the SVM trained with $K = \sum_{m=1}^M a_m K_m$.*

Choosing an interesting combination type and aggregation function is not simple because the resulting problem easily becomes either oversimplified or untractable, especially with nonlinear $\mathcal{A}$. One of the most natural examples of nonlinear outer combination type GHL corresponds to $\mathcal{A} = \max$, leading to

$$\ell_{\max}(y, t_1, \ldots, t_M) = \max_{m \in \{1,\ldots,M\}} (1 - y t_m)_+.$$

In Section 4.3, we will see that $\ell_{\max}$ has the advantage of being convex and of leading to a localized MKL solution at the same time. The *signed maximum* on $\mathbb{R}^M$ defined as $\operatorname{smax}(v) = v_{\arg\max|v|}$, or, equivalently,

$$\operatorname{smax}(v) = \begin{cases} \min v & \text{if } \min v < -|\max v|, \\ \max v & \text{if } \max v \geq |\min v|, \end{cases} \tag{4.6}$$

is an example of aggregation function that can used to define an inner combination loss. A difference between the losses based on smax and $\ell_{\max}$ lies in that using the outer combination loss with $\ell_{\max}$, we penalize each training sample according to the worse kernel, while in the inner combination loss with smax, the predictions of the kernels are aggregated independently of the target sample label, and thus independently of what the worse kernel is.

4.2.1 Representer Theorem

We extend the classical representer theorem from the single kernel case to the minimizer of (4.5). Recall that the product RKHS is $H = H_1 \times \ldots \times H_M$. In addition, we define for any kernel K_m the sample space as

$$\tilde{H}_m = \operatorname{span}\{\phi_m(x_i) : 1 \leq i \leq N\},$$

and the joint sample space as $\tilde{H} = \tilde{H}_1 \times \ldots \times \tilde{H}_M$.

Theorem 2. *The solution of (4.5) belongs to $\tilde{H}$, which is isomorphic to a subset of $\mathbb{R}^{MN}$.*

Proof. For $1 \leq m \leq M$, let $\tilde{H}_m^\perp$ be the subspace of H_m orthogonal to $\tilde{H}_m$. We also denote the product orthogonal by $\tilde{H}^\perp = \tilde{H}_1^\perp \times \ldots \times \tilde{H}_M^\perp$. Let $\omega = (\omega_1, \ldots, \omega_M)$ be a minimizer of (4.5). For any m, we can decompose $\omega_m \in H_m$ as $\omega_m = \tilde{\omega}_m + \omega_m^\perp$, with $\tilde{\omega}_m \in \tilde{H}_m$ and $\omega_m^\perp \in \tilde{H}_m^\perp$. For $1 \leq i \leq N$,

$$\langle \omega_m, \phi_m(x_i) \rangle = \langle \tilde{\omega}_m, \phi_m(x_i) \rangle + \left\langle \omega_m^\perp, \phi_m(x_i) \right\rangle = \langle \tilde{\omega}_m, \phi_m(x_i) \rangle.$$

Assume that $\omega \notin \tilde{H}$. Then, there exists m such that

$$\|\omega_m\|_2^2 = \|\tilde{\omega}_m\|_2^2 + \|\omega_m^\perp\|_2^2 > \|\tilde{\omega}_m\|_2^2.$$

As a consequence, $\tilde{\omega} = (\tilde{\omega}_1, \ldots, \tilde{\omega}_m)$ has a strictly smaller cost, which contradicts the optimality of ω. As a consequence, $\omega \in \tilde{H}$. □

This theorem generalizes the classical representer theorem for SVMs, for which $\tilde{H}$ will be a strict subset of $\mathbb{R}^{MN}$ of dimension N. The main difference with the classical representer theorem for SVMs lies in the fact that the loss is allowed to depend on an additional block

structure on the Hilbert space H. Note that this theorem does not guarantee that the solution will be localized (that is, that the solution subspace has dimension strictly greater than N). A study of representer theorems for similar losses and general regularizer functions can be found in (Argyriou et al., 2009). In this context, the block structure is derived from multi-task learning problems.

4.2.2 Universal Consistency

We now state the main result of this section, which requires the following notation. We consider here a distribution P over $X \times \{-1,1\}$, where $X \subset \mathbb{R}^d$ is a compact space. The classification and Hinge risks of a classifier $g : X \to \mathbb{R}$ are

$$R_P^{0\text{-}1}(g) = E_{(x,y)\sim P}\left[\mathbf{1}(\operatorname{sign} g(x) \neq y)\right], \quad R_P^{\text{Hinge}}(g) = E_{(x,y)\sim P}\Big[\ell_{\text{hinge}}(y, g(x))\Big].$$

Let ℓ be a GHL, and Q be a distribution on $X \times \{-1,1\}$, then the risk of $h : X \to \mathbb{R}^M$ is

$$R_Q(h) = E_{(x,y)\sim Q}\left[\ell\left(y, h_1(x), \ldots, h_M(x)\right)\right].$$

For simplicity, for $\omega \in H$, we write $R_Q(\omega) = R_Q(\langle \omega_1, \phi_1(\cdot)\rangle, \ldots, \langle \omega_M, \phi_M(\cdot)\rangle)$. The regularized risk is defined for $\omega \in H$ as

$$R_{Q,\lambda}(\omega) = R_Q(\omega) + \lambda \|\omega\|_H^2.$$

Let $\omega_{Q,\lambda}$ be the minimizer of $R_{Q,\lambda}$, and $f_{Q,\lambda} : X \to \mathbb{R}$ the corresponding classifier defined from $\omega_{Q,\lambda}$ through the link function. In this section, we analyze the regularized empirical risk minimization program

$$\min_{\omega \in H} R_{P_N,\lambda}(\omega),$$

where P_N is the empirical measure based on N observations from P, which corresponds to $C = \frac{1}{2N\lambda}$ in (4.5). The regularization parameter is allowed to depend on N, and when required, we explicitly write $\lambda_N = \lambda$. In addition, let

$$\delta_\lambda = \sqrt{\frac{2}{\lambda}}, \quad L_K = \sup_{x \in X} \max_{m \in \{1,\ldots,M\}} \sqrt{K_m(x,x)}.$$

We denote by $C(X,Z)$ the set of continuous functions from X into some space Z. A continuous kernel over X is *universal* if the corresponding RKHS is dense in $C(X,\mathbb{R})$. We say that a classifier f_N depending on N data points is *universally consistent* if $R_P^{0\text{-}1}(f_N) \to \inf R_P^{0\text{-}1}$ holds in probability for all distributions P over the data, where the infimum is over all classifiers. Let $B_H = \{\omega \in H : \|\omega\|_H \leq 1\}$ be the unit ball of H. In order to measure the complexity of $\delta_\lambda B_H$, we will use covering numbers. We denote by $\mathcal{N}(X, \|.\|, \varepsilon)$ the ε-covering number of X with respect to norm $\|.\|$, and by $H(X, \|.\|, \varepsilon) = \ln \mathcal{N}(X, \|.\|, \varepsilon)$ the corresponding metric entropy.

Theorem 3. *Let ℓ be a GHL, $(K_1, \ldots, K_M)$ be universal kernels, and for any $\varepsilon > 0$, assume that when $N \to \infty$,*

$$\lambda_N \to 0, \quad \text{and} \quad \frac{(L_\ell L_K)^2}{N\lambda_N} \sum_{m=1}^{M} H\left(\sqrt{\frac{2}{\lambda_N}}\, B_{H_m}, \|.\|_{H_m}, \frac{\varepsilon}{L_\ell L_K}\right) \to 0,$$

then, f_{P_N,λ_N} (the classifier associated to the minimizer of R_{P_N,λ_N} through the link function) is universally consistent.

In particular, if all kernels have metric entropy $H(B_m, \|.\|_{H_m}, \varepsilon) \leq \varepsilon^{-\rho}$ for some ρ, and L_K and L_ℓ are constant, then the sufficient condition is

$$\lambda_N = o(1), \quad \lambda_N = \Omega\left(\left(\frac{M}{N}\right)^{\frac{1}{1+\rho/2}}\right),$$

which requires that λ_N converges to zero at a limited rate. The result is different from consistency results for multi-class SVMs (Tewari and Bartlett, 2007; Glasmachers, 2010), as we consider multiple representations for each point and a single label, as opposed to a single representation and multiple labels. The proof proceeds in four steps, and is inspired by that of Steinwart (2005) for SVMs. We present an outline of the proof, while the details are deferred to Section 4.6. First, we show that $R_{P_N,\lambda}$ and $R_{P,\lambda}$ both attain their minimum over H in a ball of radius δ_λ.

Lemma 2 (Optimal risk is attained in a ball). *Let ℓ be a GHL, then for any Borel probability measure Q on $X \times \{-1, 1\}$ and $\lambda > 0$, there exists $\omega_{Q,\lambda} \in H$ such that $\|\omega_{Q,\lambda}\|_H \leq \delta_\lambda$ and*

$$R_{Q,\lambda}(\omega_{Q,\lambda}) = \inf_{\omega \in H} R_{Q,\lambda}(\omega).$$

Then, we show that over this ball, the unregularized risk R_{P_N} concentrates around R_P at a rate that we make explicit.

Lemma 3 (Concentration of $R_{P_N}(f_{P_N,\lambda})$). *Let ℓ be a GHL, then*

$$P_{P_N}\left(\left|R_{P_N}(\omega_{P_N,\lambda}) - R_P(\omega_{P_N,\lambda})\right| \geq \varepsilon\right) \leq 2e^{-\frac{2\varepsilon^2 N}{(L_\ell \delta_\lambda L_K)^2}} \prod_{m=1}^{M} \mathcal{N}\left(\delta_\lambda B_{H_m}, \|.\|_{H_m}, \frac{\varepsilon}{L_\ell L_K}\right).$$

Note that $R_{P_N,\lambda}$ also concentrates around $R_{P,\lambda}$. Then, we check that asymptotically and as the regularization vanishes, the minimum of $R_{P,\lambda}$ over H is close to that of R_P.

Lemma 4 (Minimum of $R_{P,\lambda}$ is close to that of R_P). *If $K_1, \ldots, K_m$ are universal and ℓ is a GHL, then*

$$\lim_{\lambda \to 0}\left[\inf_{\omega \in H} R_{P,\lambda}(\omega)\right] = \inf_h R_P(h)$$

where the second infimum is over all measurable functions $h : X \to \mathbb{R}^M$.

Finally, we use the fact that minimizing the surrogate risk R_P over all classifiers is sufficient for minimizing the classification risk. This is implied by the classical analysis of SVMs with hinge loss, and the following lemma.

Lemma 5 (R_P and R_P^{Hinge} have same minimum). *For any GHL ℓ,*

$$\inf_h R_P(h) = \inf_g R_P^{Hinge}(g).$$

In addition, for any $\delta > 0$, $h : X \to \mathbb{R}^M$ such that

$$R_P(h) \leq \inf_h R_P(h) + \delta,$$

then for θ a link function for ℓ, it holds that

$$R_P^{Hinge}(\theta \circ h) \leq \inf_g R_P^{Hinge}(g) + \delta.$$

The theorem then follows from these lemmas. Remark that we show that all GHLs have the same minimum risk over *all possible classifiers*. In light of this, and in the situation where one would be able to compute (and represent in a machine) the optimal classifier associated to, say, R_P^{Hinge}, there would be no interest in using GHLs. However, this is not the case, and as we will see in the next section, the advantage of using (tractable) GHLs is that it may allow to optimize the risk over a wider class of decision functions than with SVMs or linear MKL.

4.3 ℓ_p-norm Aggregation of Hinge Losses

So far, we have shown that GHLs can lead to universal consistency, and benefit from a representer theorem in $\mathbb{R}^{MN}$, but they do not in general lead to a localized solution, that is, α can be of low rank. In this section, we propose a family of losses which lead to a localized solution, based on outer aggregation with ℓ_p-norms for $p \in [1,\infty]$, also referred to as *power-means*. More precisely, we consider

$$\ell_{\text{agg},p}(y,t_1,\ldots,t_M) = M^{-1/p}\left[\sum_{m=1}^{M}(1-yt_m)_+^p\right]^{1/p} = M^{-1/p}\left\|((1-yt_m)_+)_{m\in[M]}\right\|_p$$

for $p \in [1,\infty]$, which is a convex GHL.

4.3.1 Dual Problem

We now go into more detail into the optimization problem associated with these losses, and derive the corresponding dual problem. Due to strong duality, optimizing the dual problem will be equivalent. Although we have not included a bias term so far for convenience of the theoretical analysis (see also (Steinwart et al., 2011)), our formulation will include one offset per kernel, such that the classifier is

$$f(x) = \theta\left(\langle\omega_1,\phi_1(x)\rangle + b_1,\ldots,\langle\omega_M,\phi_M(x)\rangle + b_M\right).$$

The methods that we propose correspond to the primal problem

$$\min_{\substack{\omega=(\omega_1,\ldots,\omega_M)\in H\\ b\in\mathbb{R}^M}}\left[\frac{1}{2}\sum_{m=1}^{M}\|\omega_m\|_2^2 + C\sum_{i=1}^{N}\ell_{\text{agg},p}\left(y_i,\langle\omega_1,\phi_1(x_i)\rangle + b_1,\ldots,\langle\omega_M,\phi_M(x_i)\rangle + b_M\right)\right].$$

The offsets are not regularized, as is common for SVMs. We omit in the following the scaling factor $M^{1/p}$ as this can be integrated into C. Using slack variables, the problem is

$$\inf_{\substack{\omega\in H\\ \xi\in\mathbb{R}^{N\times M}\\ b\in\mathbb{R}^M}}\left[\frac{\|\omega\|_H^2}{2} + C\sum_{i=1}^{N}\|\xi_{i,\cdot}\|_p\right]\text{s.t.}\ \begin{array}{l}\xi_{i,m}\geq 0 \text{ for } i\in[N], m\in[M]\\ \xi_{i,m}\geq 1-y_i\left[\langle\omega_m,\phi_m(x_i)\rangle + b_m\right] \text{ for } i\in[N],\, m\in[M].\end{array}$$

Introducing Lagrange multipliers $\alpha,\beta\in\mathbb{R}_+^{N\times M}$, the Lagrangian is

$$L = \frac{\|\omega\|_H^2}{2} + \sum_{i=1}^{N}\left[C\|\xi_{i,\cdot}\|_p - \langle\xi_{i,\cdot},\alpha_{i,\cdot}+\beta_{i,\cdot}\rangle\right] + \sum_{i=1}^{N}\sum_{m=1}^{M}\alpha_{i,m}\left(1-y_i\left[\langle\omega_m,\phi_m(x_i)\rangle + b_m\right]\right).$$

Hence,

$$\begin{aligned}
&\inf_{\substack{\omega\in H\\ \xi\in\mathbb{R}^{N\times M}\\ b\in\mathbb{R}^M}} L(\alpha,\beta,\omega,\xi,b)\\
&\quad= \sum_{i=1}^{N}\inf_{\xi_{i,\cdot}\in\mathbb{R}^N}\left[C\,\|\xi_{i,\cdot}\|_p - \langle\xi_{i,\cdot},\alpha_{i,\cdot}+\beta_{i,\cdot}\rangle\right]\\
&\qquad+\inf_{\substack{\omega\in H\\ b\in\mathbb{R}^M}}\left[\frac{\|\omega\|_H^2}{2}+\sum_{i=1}^{N}\sum_{m=1}^{M}\alpha_{i,m}\left(1-y_i\left[\langle\omega_m,\phi_m(x_i)\rangle+b_m\right]\right)\right]\\
&\quad= -C\sum_{i=1}^{N}\sup_{\xi_{i,\cdot}\in\mathbb{R}^M}\left[\left\langle\xi_{i,\cdot},\frac{\alpha_{i,\cdot}+\beta_{i,\cdot}}{C}\right\rangle-\|\xi_{i,\cdot}\|_p\right]\\
&\qquad+\inf_{\substack{\omega\in H\\ b\in\mathbb{R}^M}}\left[\frac{\|\omega\|_H^2}{2}+\sum_{i=1}^{N}\sum_{m=1}^{M}\alpha_{i,m}\left(1-y_i\left[\langle\omega_m,\phi_m(x_i)\rangle+b_m\right]\right)\right].
\end{aligned}$$

Denote by f^* the convex conjugate of a function $f:\mathbb{R}^M\to\mathbb{R}$, such that

$$f^*(z)=\sup_{x\in\mathbb{R}^M}\left[\langle x,z\rangle-f(x)\right].$$

For any $i\in[N]$, the value of the maximization problem over $\xi_{i,\cdot}$ is precisely the convex conjugate of the ℓ_p-norm as a function of $\mathbb{R}^M$. As is classical, the convex conjugate of a norm $\|\cdot\|$ over $\mathbb{R}^M$ is the indicator function of the unit ball of the so-called dual norm $\|\cdot\|_*$ defined for $z\in\mathbb{R}^M$ as $\|z\|_*=\sup_{x\in\mathbb{R}^M}\{\langle x,z\rangle:\|x\|\le 1\}$. Formally,

$$(\|\cdot\|)^*=\delta_{\|\cdot\|_*\le 1},$$

where δ_A is such that $\delta_A(x)=0$ if $x\in A$, and $\delta_A(x)=\infty$ otherwise. The dual norm of the ℓ_p-norm is well known to be the ℓ_q-norm, for q such that $\frac{1}{p}+\frac{1}{q}=1$. As a consequence, for $C\ge 0$,

$$-C\sum_{i=1}^{N}\sup_{\xi_{i,\cdot}\in\mathbb{R}^M}\left[\left\langle\xi_{i,\cdot},\frac{\alpha_{i,\cdot}+\beta_{i,\cdot}}{C}\right\rangle-\|\xi_{i,\cdot}\|_p\right]=\begin{cases}0 & \text{if } \left\|\alpha_{i,\cdot}+\beta_{i,\cdot}\right\|_q\le C \text{ for } i\in[N],\\ -\infty & \text{otherwise.}\end{cases}$$

As β has no influence on the second term, the condition can be equivalently written $\|\alpha_{i,\cdot}\|_q\le C$ for $i\in[N]$. In the second term, minimizing with respect to $b=(b_1,\dots,b_M)$ leads to the condition $\sum_{i=1}^N y_i\alpha_{i,m}=0$ for all $m\in[M]$ (which we abbreviate in the following by $\alpha\in\mathbf{B}$), while minimizing with respect to ω gives $\omega_m=\sum_{i=1}^N\alpha_{i,m}y_i\phi_m(x_i)$. This leads to the dual problem

$$\sup_{\substack{\alpha\ge 0\\ \beta\ge 0}}\inf_{\substack{\omega\in H\\ \xi\in\mathbb{R}^{N\times M}\\ b\in\mathbb{R}^M}} L=\sup_{\alpha\ge 0}\left[\mathbf{1}^T\alpha\,\mathbf{1}-\frac{1}{2}\sum_{m=1}^{M}\alpha_{\cdot,m}^T H_m\alpha_{\cdot,m}\right]\text{ s.t. }\|\alpha_{i,\cdot}\|_q\le C\text{ for }i\in[N],\quad\alpha\in\mathbf{B},$$

with, as previously, $H_m=\operatorname{diag}(y)K_m\operatorname{diag}(y)$. The quadratic part can be rewritten as $\operatorname{vec}(\alpha)^T Q\operatorname{vec}(\alpha)$ with vec the concatenation operator, and

$$Q=\begin{pmatrix}H_1 & 0 & 0\\ 0 & \ddots & 0\\ 0 & 0 & H_M\end{pmatrix}.$$

In particular, this is a convex quadratic program. From this dual problem, it follows that the decision function is localized, and of the form

$$f(x) = \theta\left(\sum_{i=1}^{N} y_i \alpha_{i,1} K_1(x_i, x) + b_1, \ldots, \sum_{i=1}^{N} y_i \alpha_{i,M} K_M(x_i, x) + b_M, \right).$$

As already mentioned, this corresponds to a slight generalization to (4.2). The biases can be recovered as follows: let $\mathcal{D}$ denote the previous dual objective, such that the KKT optimality conditions at an optimal dual estimate α^* imply that, for any $m \in [M]$,

$$\max_{i \in I_u(m)} y_i \nabla \mathcal{D}(\alpha^*) \leq b_m \leq \min_{i \in I_d(m)} y_i \nabla \mathcal{D}(\alpha^*), \tag{4.7}$$

where for $m \in [M]$,

$$I_u(m) = \left\{i : y_i = 1, \quad \|\alpha^*_{i,\cdot}\|_q < C\right\} \cup \left\{i : y_i = -1, \quad \alpha^*_{i,m} > 0\right\},$$
$$I_d(m) = \left\{i : y_i = 1, \quad \alpha^*_{i,m} > 0\right\} \cup \left\{i : y_i = -1, \quad \|\alpha^*_{i,\cdot}\|_q < C\right\}.$$

As is common, we use for numerical stability the average over all $i \in I_u(m) \cap I_d(m)$. When this set is empty, b_m may not be unique, and we set b_m to the midpoint of the interval defined by Equation (4.7).

In the presence of the ℓ_q-norm constraints, the dual problem can be formulated as a second-order cone programming (SOCP) problem, which consists in minimizing a linear function over the intersection of second order cones of the form

$$\|A_i x + b_i\|_2 \leq \langle c_i, x \rangle + d_i.$$

As shown in Alizadeh and Goldfarb (2003), ℓ_q-norm constraints with $1 < q < \infty$ can be reformulated using such second order cone constraints, although this transformation results in a very significant increase in the dimensionality of the problem. This simplifies to a QP when $p \in \{1, \infty\}$, or a quadratically-constrained quadratic program (QCQP) for $p = 2$. When N and M are moderate, the estimator can hence be obtained using off-the-shelf optimization packages. Naturally, the case where $p = 1$ is readily solved using any SVM solver.

4.3.2 Regularization

The dual is regularized through a $\|\cdot\|_{\infty,q}$-norm constraint, where for $\alpha \in \mathbb{R}^{N \times M}$,

$$\|\alpha\|_{\infty,q} = \max_{i \in [N]} \|\alpha_{i,\cdot}\|_q,$$

which is dual to the $\|\cdot\|_{1,p}$-norm that we use to aggregate the losses corresponding to different kernels. This latter norm is often used to induce group sparsity - see, e.g., (Sra, 2011) and references therein. We do not require any additional regularization, which has to be contrasted with classical method for even non-localized MKL. In the early days of MKL, the interest was focused on ℓ_1 regularization of the kernel weights, i.e., $\Omega(s) = \lambda \|s\|_1$, or on constrained versions of this, such as in SimpleMKL (Rakotomamonjy et al., 2008). However, this type of regularization leads to a non-differentiable dual problem, which

precludes coordinate descent methods such as SMO that have been most influential in terms of accessibility and scalability of SVMs. More recently, there has been interest in ℓ_p regularization for $p > 1$, i.e., $\Omega(s) = \lambda \|s\|_p^p$, or constrained versions of this (Kloft et al., 2009). Although this produces non-sparse mixtures of kernels, this has been shown to yield at least comparable classification performance as ℓ_1 regularization in a variety of settings. In many penalized variants of non-localized MKL, the amount of regularization has to be adjusted in addition to the C parameter, which can be quite expensive. For $p > 1$, this was alleviated by Kloft et al. (2009), who have studied the constrained version where $\|s\|_p \leq 1$, and shown that there exists choices of C such that the problem is equivalent to the (C, λ)-penalized problem. Along the lines of these successive developments for MKL, our method based on ℓ_p-norm aggregation of hinge losses can be generalized to arbitrary aggregation functions $\mathcal{A}$, which we believe could lead to some other interesting formulations. For a GHL of outer combination type with aggregation function $\mathcal{A}$, one similarly obtains the dual problem

$$\sup_{\alpha \geq 0} \left[\mathbf{1}^T \alpha \, \mathbf{1} - \frac{1}{2} \sum_{m=1}^{M} \alpha_{\cdot,m}^T H_m \alpha_{\cdot,m} - C \sum_{i=1}^{N} \mathcal{A}^*(\alpha_{i,\cdot}) \right] \text{ s.t. } \alpha \in \mathbf{B},$$

where $\mathcal{A}^* : \mathbb{R}^M \to \mathbb{R}$ is the convex conjugate of $\mathcal{A}$, such that aggregation functions may be designed to enforce certain penalizations, while retaining a single regularization parameter.

4.3.3 Link Functions

In order to define classifiers, we need to specify link functions. As $\|x\|_\infty \leq \|x\|_p \leq \|x\|_1$ for $p \in [1, \infty]$ and $x \in \mathbb{R}^M$, any link function for $p = \infty$ can be rescaled to obtain a link function for any other value of p. Note that ultimately, only the sign of the link function is used for testing. The mean of the kernel scores provides such a valid link function.

Lemma 6. *$\mathtt{avg}(t_1, \ldots, t_M) = \frac{1}{M} \sum_{m=1}^{M} t_m$ is a link function for $\ell_{agg,\infty}$, i.e., for any y and $t \in \mathbb{R}^M$,*

$$\ell_{agg,\infty}(y, t_1, \ldots, t_M) \geq (1 - y\,\mathtt{avg}(t_1, \ldots, t_M))_+ .$$

4.4 Experiments

In this section, we assess on real data the performance of the two proposed methods, together with that of alternative methods for localized MKL. We consider the following baselines in our experiments:

- **ℓ_p-MKL.** We used linear MKL, both with ℓ_1 and ℓ_2 regularization, using the `SHOGUN` library. For ℓ_1 regularization, we used the SILP formulation and chunking-based algorithm from Sonnenburg et al. (2006), where the nonnegative weights of the kernels are constrained to sum to one, while the ℓ_p-constrained formulation and algorithm from Kloft et al. (2009) was used for ℓ_2 regularization.

- **HFM.** We implemented Heterogeneous Feature Machines from Cao et al. (2009) which uses a logistic loss with a $\ell_{1,2}$ regularization, as described in (4.4).

- **LMKL.** We used the localized MKL formulation and implementation introduced by Gönen and Alpaydin (2008). The gating functions are based on the original vector data representation when available, and the method was omitted from the comparisons when unavailable.

- **PS-MKL.** We implemented PS-MKL as described in (Yang et al., 2010) using alternate optimization. When $(\beta_{i,m})$ is fixed, (α_i) is optimized using `LIBSVM`. When (α_i) is fixed, $(\beta_{i,m})$ is optimized using a linear programming solver. Note that with respect to $(\beta_{i,m})$, the problem is actually a semi-infinite linear program (SILP), which can be solved by constraint generation. As described in the original paper, we add a constraint at every iteration, such that the cost of solving the corresponding linear program (in dimension MN) increases with the iteration number.

To the best of our knowledge, there was no previous experimental comparison of any two of HFM, LMKL, and PS-MKL, although they address a similar question. In addition, we consider the following algorithms based on ℓ_p-norm hinge loss aggregation:

- $\ell_{\mathbf{agg},1}$ **- average of SVMs.** This corresponds to $p = 1$, and amounts to aggregating the decision of independent SVMs.

- $\ell_{\mathbf{agg},\infty}$ **- maximum of hinge losses.** This corresponds to $p = \infty$, and amounts to aggregation of the hinge losses using the max, or, alternatively, to a $\|.\|_{\infty,1}$ constraint in the dual.

For each of these two methods that we propose, we use `avg` as link function.

4.4.1 UCI Datasets

We conducted experiments on `wpbc`, `ionosphere`, `sonar`, `liver` and `pima` (see Table 4.1). In each case, we used 25 Gaussian kernels with bandwidths regularly distributed in logarithmic space. The SVM parameter C is selected using 10-fold cross validation. The average testing accuracy (in percentage) over 10 train/test splits is shown in Figure 4.1.

Dataset	Num. training examples	Num. testing examples	Num. features
wpbc	136	58	33
ionosphere	246	105	34
sonar	146	62	60
liver	140	201	6
pima	538	230	8

Table 4.1: Characteristics of UCI datasets used

We first note that in `pima` and `sonar`, MKL performed as well or nearly as well as the localized MKL methods. This has also been observed in (Gönen and Alpaydın, 2011) when comparing LMKL to various MKL variants, and is likely a consequence of the important expressive power of the Gaussian kernels. No method for localized MKL seem to exhibit uniformly better results. Among localized methods, LMKL, HFM, and $\ell_{\text{agg},1}$ tend to perform best.

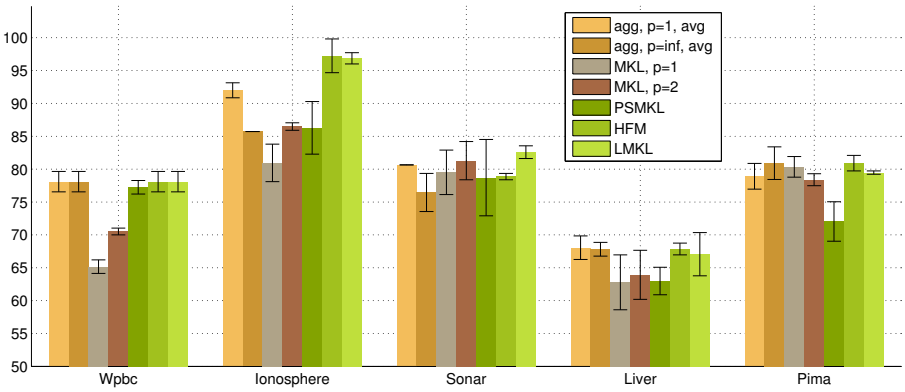

Figure 4.1: Testing accuracy (in percent) on UCI datasets

4.4.2 Image Classification

In addition, we performed experiments on classical image classification tasks:

- **`Caltech101`.** This dataset (Fei-Fei et al., 2007) comprises images from 101 classes, with about 40 to 800 images per category. We used exponential kernels based on χ_2 divergences over PHOW gray/color and SSIM features at three different scales, following Vedaldi et al. (2009), and amounting to a total of 9 kernels.

- **`Oxford Flowers`.** This dataset consists in images from 102 classes of flowers (Nilsback and Zisserman, 2008), with about 40 to 250 images per category, and significant pose, scale and illumination variations. We used kernels from Nilsback and Zisserman (2008), which consists in exponential kernels based on χ_2 divergences over color histograms, SIFT features both on pre-segmented foreground region and on the corresponding boundary, and histogram of gradients (HOG) features, resulting in a total of 4 kernels.

In each case, we used three training/testing splits, with 15 positive examples per class. One-vs-rest (OVR) classification was used, and all algorithms were modified to include different weights C_+ and C_- in their loss for positive and negative examples, respectively. Although oversampling positive examples can sometimes be beneficial (Perronnin et al., 2012), we used unbiased weights $C_+ = C/N_+, C_- = C/N_-$, where N_+ and N_- are the number of positive and negative examples in the training set, respectively. The same parameter C was used for all classes. The results are summarized in Figure 4.2.

The methods that we propose achieve the highest testing accuracies in both benchmarks. We obtain similar accuracies both for $p = 1$ and $p = \infty$. This is discussed further in Section 4.5. On the other hand, we obtained the lowest test accuracy using PS-MKL. This is surprising, as particularly good results compared to various MKLs are provided on this dataset in (Yang et al., 2010). First note that we use different and somewhat simpler kernels here, such that we do not expect to necessarily match or exceed accuracies provided in the original paper. More importantly, PS-MKL sometimes failed to converge within our maximum time window of two hours for some combinations of parameters and classes, which can have a large impact on OVR classification. Although this could be

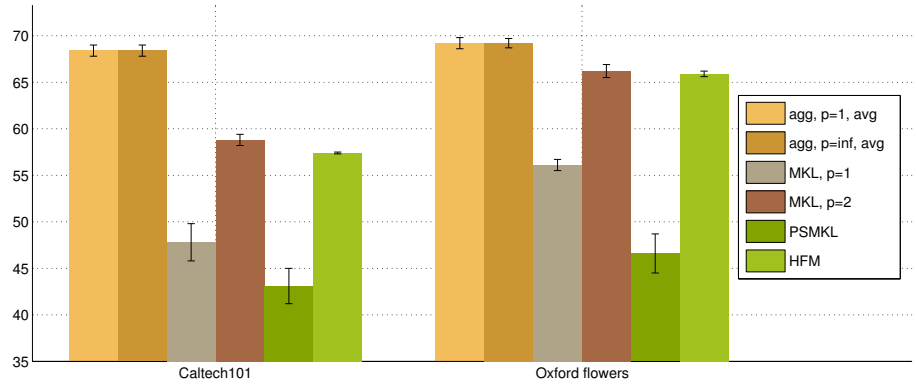

Figure 4.2: Testing accuracy (in percent) on Caltech101 and Oxford flowers datasets

fixed in principle (e.g., using multiple initializations), we believe that this is an inherent drawback of alternate optimization.

In the previous experiments, we do not provide formal timings for training, as there are no publicly available implementations of either HFM and PS-MKL, and not all algorithms have a computational cost that can be easily summarized in terms of number of calls to a SVM solver or gradient evaluations as is often done. Informally, the two variants of MKL, as well as LMKL and our method with $\ell_{\text{agg},1}$ were the fastest, followed by HFM and our method with $\ell_{\text{agg},\infty}$, while PS-MKL was the slowest.

4.5 Discussion

The first method that we propose based on averaging decision functions from independent SVMs clearly achieves the best tradeoff between accuracy and training time (as well as ease of implementation). The good performance of this simple method may be informally explained in the following way: averaging decisions associated to different kernels amounts to a weighted voting. Due to the large margin property of SVMs, kernels that are unsure are likely to abstain by producing low magnitude decisions, while kernels that are confident are likely to sway the final decision.

In all our results, $p = 1$ is always at least as good as $p = \infty$. Although we have not reported such results here, the methods corresponding to aggregation with $p \in]1, \infty[$ are also outperformed by $p = 1$ on our experiments. Hence, coupling the kernels does not in general seem to improve accuracy. This can be explained through the idea that by forcing kernels to agree through the coupling, the classifiers associated to each kernel are individually not as good as without coupling. The previous observations reflects a result from MKL (Gehler and Nowozin, 2009), that under reasonable scalings of the kernels, using a SVMs with the average kernel $\bar{K} = \frac{1}{M} \sum_{m=1}^{M} K_m$ leads in many situation to a performance close to that of full-fledged MKL, which is very similar to the phenomenon that we observe here.

4.6 Proofs

4.6.1 Proof of Lemma 2

Proof. Let $\varepsilon \in (0,1)$, and $\omega_\varepsilon \in H$ that is ε-optimal for the risk $R_{Q,\lambda}$, such that

$$R_{Q,\lambda}(\omega_\varepsilon) \leq \inf_{\omega \in H} R_{Q,\lambda}(\omega) + \varepsilon.$$

Since

$$\lambda \|\omega_\varepsilon\|_H^2 \leq R_{Q,\lambda}(\omega_\varepsilon) \leq R_{Q,\lambda}(0) + \varepsilon \leq 1 + \varepsilon \leq 2,$$

such that $\|\omega_\varepsilon\|_H \leq \delta_\lambda$. The rest of the proof proceeds as in (Steinwart, 2005, Lemma 3.1) for SVMs. □

4.6.2 Proof of Lemma 3

Proof. We will consider covering numbers with respect to

$$\|\omega\|_{H,\infty} = \max_{m \in [M]} \|\omega_m\|_{H_m}.$$

Let $(\omega^{(1)}, \ldots, \omega^{(n)}) \in H^n$ be an ε-cover of $\delta_\lambda B_H$ with respect to $\|.\|_{H,\infty}$, and for any $i \leq n$, decompose $\omega^{(i)} = (\omega_1^{(i)}, \ldots, \omega_M^{(i)})$. In particular, for any $\omega = (\omega_1, \ldots, \omega_M) \in \delta_\lambda B_H$, there exists $i \leq n$ such that $\|\omega - \omega^{(i)}\|_{H,\infty} < \varepsilon$. Let $x \in X$, and $y \in \{-1,1\}$ be fixed. For any m,

$$\left|\langle \omega_m - \omega_m^{(i)}, \phi_m(x) \rangle\right| \leq \|\omega_m - \omega_m^{(i)}\|_{H_m} \|\phi_m(x)\|_{H_m} \leq \varepsilon \sqrt{K_m(x,x)}.$$

As ℓ is a GHL,

$$\begin{aligned}
&\left|\ell\left(y, \langle \omega_1, \phi_1(x)\rangle, \ldots, \langle \omega_M, \phi_M(x)\rangle\right) - \ell\left(y, \langle \omega_1^{(i)}, \phi_1(x)\rangle, \ldots, \langle \omega_M^{(i)}, \phi_M(x)\rangle\right)\right| \\
&\quad \leq L_\ell \max_{m \in [M]} \left|\langle \omega_m - \omega_m^{(i)}, \phi_m(x)\rangle\right| \\
&\quad \leq \varepsilon L_\ell \max_{m \in [M]} \sqrt{K_m(x,x)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
&\sup_{\substack{x \in X, \\ y \in \{-1,1\}}} \left|\ell\left(y, \langle \omega_1, \phi_1(x)\rangle, \ldots, \langle \omega_M, \phi_M(x)\rangle\right) - \ell\left(y, \langle \omega_1^{(i)}, \phi_1(x)\rangle, \ldots, \langle \omega_M^{(i)}, \phi_M(x)\rangle\right)\right| \\
&\qquad \leq \varepsilon L_\ell \left[\sup_{x \in X} \max_{m \in \{1,\ldots,M\}} \sqrt{K_m(x,x)}\right].
\end{aligned}$$

As a consequence, the functions $f_i = \ell\left(\cdot, (\langle \omega_m^{(i)}, \phi_m(\cdot)\rangle)_{m \in [M]}\right)$, $1 \leq i \leq n$ form an $(\varepsilon L_\ell L_K)$-covering of $\mathcal{F} = \left\{\ell\left(\cdot, (\langle \omega_m, \phi_m(\cdot)\rangle)_{m \in [M]}\right) : \omega \in \delta_\lambda B_H\right\}$, which is a subset of the space of continuous functions from $X \times \{-1,1\}$ to $\mathbb{R}_+$ (endowed with the supremum norm). Hence,

$$\mathcal{N}(\mathcal{F}, \|.\|_\infty, \varepsilon) \leq \mathcal{N}\left(\delta_\lambda B_H, \|.\|_{H,\infty}, \frac{\varepsilon}{L_\ell L_K}.\right)$$

For any $f \in \mathcal{F}$, we have $\|f\|_\infty \leq 1 + L_\ell \, \delta_\lambda \, L_K$. Hence, using Hoeffding's inequality,

$$P_{P_N}\left(\sup_{\omega \in \delta_\lambda B_H} \left| R_{P_N}(\omega) - R_P(\omega) \right| \geq \varepsilon \right) \leq 2\mathcal{N}\left(\delta_\lambda B_H, \|.\|_{H,\infty}, \frac{\varepsilon}{L_\ell L_K} \right) e^{-\frac{2\varepsilon^2 N}{(L_\ell \, \delta_\lambda \, L_K)^2}}.$$

Finally, for any $\bar{\varepsilon} > 0$,

$$\mathcal{N}(\delta_\lambda B_H, \|.\|_{H,\infty}, \bar{\varepsilon}) \leq \prod_{m=1}^{M} \mathcal{N}(\delta_\lambda B_{H_m}, \|.\|_{H_m}, \bar{\varepsilon}).$$

The result follows. □

4.6.3 Proof of Lemma 4

Proof. Let $\varepsilon > 0$ and $\omega_\varepsilon \in H$ such that $R_P(\omega_\varepsilon) \leq \inf_{\omega \in H} R_P(\omega) + \frac{\varepsilon}{2}$. For all $\lambda \leq \frac{\varepsilon}{2\|\omega_\varepsilon\|_H^2}$,

$$\inf_{\omega \in H} R_{P,\lambda}(\omega) \leq R_{P,\lambda}(\omega_\varepsilon) \leq R_P(\omega_\varepsilon) + \frac{\varepsilon}{2} \leq \inf_{\omega \in H} R_P(\omega) + \varepsilon.$$

Hence,

$$\lim_{\lambda \to 0} \left[\inf_{\omega \in H} R_{P,\lambda}(\omega) \right] = \inf_{\omega \in H} R_P(\omega).$$

We sketch the rest of the proof, which consists in showing that one can approximate the optimal risk over all measurable functions from X to $\mathbb{R}^M$ using bounded functions, and in turn that bounded functions from X to $\mathbb{R}^M$ can be approximated by elements from the RKHS. Let $g = (g_1, \ldots, g_M)$ with $\|g_m\|_\infty < \infty$ for $m \in [M]$. From classical arguments, there exists (g^n) such that $g^n = (g_1^n, \ldots, g_M^n) \in C(X, \mathbb{R}^M)$, with $\sup_{m \in [M]} \|g_m^n\|_\infty \leq \sup_{m \in [M]} \|g_m\|_\infty$, and $\sup_{m \in [M]} \|g_m^n - g_m\|_1 \to 0$. For any m, since K_m is universal, H_m is dense in $C(X, \mathbb{R})$. As the closure of a cartesian product of sets coincides with the cartesian product of the closures, H is dense in $C(X, \mathbb{R}) \times \ldots \times C(X, \mathbb{R}) = C(X, \mathbb{R}^M)$. As a consequence, there exists (h^n) such that $h^n = (h_1^n, \ldots, h_M^n) \in H$, and $\sup_{m \in [M]} \|h_m^n - g_m^n\|_\infty \to 0$. Hence, for n large enough, h^n is bounded and the sequence converges point wise P_X-almost everywhere to g. The results follows from the fact the a generalized hinge loss is a Lipschitz function, and Lebesgue's theorem, as in (Steinwart and Christmann, 2008, Proposition 5.27). □

4.6.4 Proof of Lemma 5

Proof. Let $g_0 : X \to \mathbb{R}$ such that $R_P^{\text{Hinge}}(g_0) = \inf_g R_P^{\text{Hinge}}(g)$. Consider $h_0 : X \to \mathbb{R}^M$, such that $h_0(x) = (g_0(x), \ldots, g_0(x))$. Then,

$$R_P(h_0) = R_P^{\text{Hinge}}(g_0) \leq \inf_g R_P^{\text{Hinge}}(g).$$

As a consequence, $\inf_h R_P(h) \leq \inf_g R_P^{\text{Hinge}}(g)$. Conversely, let $h_0 : X \to \mathbb{R}^M$, such that $h_0 = (h_0^1, \ldots, h_0^M)$, and let $\psi(x) = \theta(h_0^1(x), \ldots, h_0^M(x))$, such that

$$R_P(h_0) \geq R_P^{\text{Hinge}}(\psi) \geq \inf_g R_P^{\text{Hinge}}(g).$$

As a consequence, $\inf_h R_P(h) \geq \inf_g R_P^{\text{Hinge}}(g)$. This shows that

$$\inf_h R_P(h) = \inf_g R_P^{\text{Hinge}}(g).$$

□

4.6.5 Proof of Theorem 3

Proof. Let $\epsilon > 0$. From (Steinwart, 2005, Proposition 3.3), there exist $\delta > 0$, such that for all measurable $g_0 : X \times \{-1, 1\} \to \mathbb{R}$,

$$R_P^{\text{Hinge}}(g_0) \leq \inf R_P^{\text{Hinge}} + \delta \Rightarrow R_P^{0\text{-}1}(g_0) \leq R_P^{0\text{-}1} + \epsilon.$$

Let us pick such a $\delta > 0$. Since $\lambda \to 0$, by Lemma 4, we have $\lim_{N\to\infty} \left[\inf_{\omega\in H} R_{T,\lambda_N}(\omega)\right] = \inf_h R_P(h)$, and as a consequence, there exists N_1 such that for $N \geq N_1$,

$$\left|\inf_{\omega\in H} R_{T,\lambda_N}(\omega) - \inf_h R_P(h)\right| \leq \frac{\delta}{3}. \tag{4.8}$$

Since $\lambda \to 0$, there exists $\rho > 0$ and $N_2 \in \mathbb{N}$ such that for $N \geq N_2$ we have by Lemma 3 and using the condition on (λ_N),

$$\left|R_{P_N}(\omega_{P_N,\lambda}) - R_P(\omega_{P_N,\lambda})\right| < \frac{\delta}{3} \tag{4.9}$$

occurs with probability less than $\varepsilon/2$. Similarly, using Hoeffding's inequality,

$$\left|R_{P_N}(\omega_{P,\lambda}) - R_P(\omega_{P,\lambda})\right| < \frac{\delta}{3} \tag{4.10}$$

occurs with probability less than $\varepsilon/2$. Hence, the probability that both of these two hold is at least $1 - \varepsilon$. For $N \geq \max(N_1, N_2)$, with probability $1 - \epsilon$,

$$\begin{aligned}
R_P(\omega_{P_N,\lambda_N}) &\leq R_{P,\lambda_N}(\omega_{P_N,\lambda_N}) \\
&\leq R_{P_N,\lambda_N}(\omega_{P_N,\lambda_N}) + \frac{\delta}{3} \text{ by (4.9)} \\
&\leq R_{P_N,\lambda_N}(\omega_{P,\lambda_N}) + \frac{\delta}{3} \text{ by definition of } \omega_{T,\lambda_N} \\
&\leq R_{P,\lambda_N}(\omega_{P,\lambda_N}) + \frac{2\delta}{3} \text{ by (4.10)} \\
&\leq \inf_h R_P(h) + \delta \text{ by (4.8)}\,.
\end{aligned}$$

Finally, using Lemma 5, with $h : X \to \mathbb{R}^M$ associated to ω_{P_N,λ_N} and $g = \theta \circ h$, it holds that

$$R_P^{\text{Hinge}}(g) \leq \inf_g R_P^{\text{Hinge}}(g) + \delta.$$

Then, $R_P^{\text{Hinge}}(g) \leq \inf R_P^{\text{Hinge}} + \delta$, and as a consequence, $R_P^{0\text{-}1}(\operatorname{sign} g) \leq R_P^{0\text{-}1} + \epsilon$. □

Part II

Detection of Structured Objects

5

Detection of Correlations with Adaptive Sensing

"Half the money I spend on advertising is wasted; the trouble is, I don't know which half."

— John Wanamaker (attributed)

Contents

5.1 Introduction . . . 104
5.1.1 Model . . . 104
5.1.2 Adaptive vs. Non-adaptive Sensing and Testing . . . 105
5.1.3 Uniform Sensing and Testing . . . 107
5.1.4 Related Work . . . 108
5.1.5 Outline . . . 109
5.1.6 Notation . . . 109
5.2 Lower Bounds . . . 110
5.3 Adaptive Tests . . . 113
5.3.1 Sequential Thresholding . . . 113
5.3.2 The Case of k-intervals . . . 115
5.3.3 The Case of k-sets: Randomized Subsampling . . . 118
5.4 Unnormalized Correlation Model . . . 120
5.4.1 Model and Extensions of Previous Results . . . 120
5.4.2 The Case of k-sets . . . 120
5.5 Discussion . . . 122
5.6 Proofs . . . 124
5.6.1 Inequalities and KL Divergences . . . 124
5.6.2 Proof of Bound on KL Divergence . . . 125
5.6.3 Proof of Proposition 18 . . . 125
5.6.4 Proof of Proposition 19 . . . 126
5.7 Appendix: Extensions to Unnormalized Model . . . 127
5.7.1 Uniform (non-adaptive) Lower Bound . . . 127
5.7.2 Uniform (non-adaptive) Upper Bound . . . 127
5.7.3 KL Divergences . . . 127

5.1 Introduction

In this chapter, we are interested in the following statistical problem: given multiple observations from a Gaussian multivariate distribution we want to test whether the corresponding covariance matrix is diagonal against non-diagonal alternatives. This type of problem has recently received a lot of attention in the literature, where different models and choices of non-diagonal covariance alternatives were considered (Hero and Rajaratnam, 2012; Arias-Castro et al., 2014, 2012; Berthet and Rigollet, 2013; Cai et al., 2013). In this work, we consider the detection of sparse *positive* correlations, which has been treated in the case of a unique multivariate sample (Arias-Castro et al., 2014), or of multiple samples (Arias-Castro et al., 2012). However, our work deviates from the existing literature in that we consider an *adaptive sensing* or *sequential experimental design* setting. More precisely, data is collected in a sequential and adaptive way, where data collected at earlier stages informs the collection of data in future stages. Adaptive sensing has been studied in the context of other detection and estimation problems, such as in detection of a shift in the mean of a Gaussian vector (Castro, 2012; Haupt et al., 2009), in compressed sensing (Arias-Castro et al., 2013; Haupt et al., 2012; Castro, 2012), in experimental design, optimization with Gaussian processes (Srinivas et al., 2010), and in active learning (Chen and Krause, 2013). Adaptive sensing procedures are quite flexible, as the data collection procedure can be "steered" to ensure most collected data provides important information. As a consequence, procedures based on adaptive sensing are often associated with better detection or estimation performances than those based on non-adaptive sensing with a similar measurement budget. In this work, our objective is to determine whether this is also the case for detection of sparse positive correlations, and if so, to quantify how much can be gained.

5.1.1 Model

Let $U^t \in \mathbb{R}^n$, $t = 1, 2, \ldots$ be independent and identically distributed (i.i.d.) normal random vectors with zero mean and covariance matrix Σ_S, where S is a subset of $\{1, \ldots, n\}$. Let $\rho > 0$ and define the covariance matrix as

$$(\Sigma_S)_{i,j} = \begin{cases} 1, & i = j \\ \rho, & i \neq j, \text{ with } i, j \in S \\ 0, & \text{otherwise} \end{cases} .$$

Our main goal is to solve the hypothesis testing problem

$$H_0 : S = \emptyset$$
$$H_1 : S \in \mathcal{C} ,$$

where $\mathcal{C}$ is some class of non-empty subsets of $\{1, \ldots, n\}$, each of size k. In other words, under the alternative hypothesis, there exists an unknown subset $S \in \mathcal{C}$ such that corresponding components are positively correlated with strength $\rho > 0$. We will often denote the elements of S as the subset of *contaminated* coordinates. In all cases we assume that the cardinality of each $S \in \mathcal{C}$ is the same: $\mathrm{card}(S) = k$. We consider the following types of classes $\mathcal{C}$ for the contaminated coordinates:

- **k-intervals**: all sets of k contiguous coordinates, of the form $\{z, z+1, \ldots, z+k-1\}$ for some $1 \leq z \leq n-k+1$; this class has size linear in n, and we denote it by $\mathcal{C}_{[k]}$.

- **disjoint k-intervals**: the class $\mathcal{D}_{[k]}$ defined as

$$\mathcal{D}_{[k]} = \{I_1, \ldots, I_{\lfloor n/k \rfloor}\}, \quad I_j = \{(j-1)k+1, \ldots, jk\},\ j \in \{1, \ldots, \lfloor n/k \rfloor\}.$$

- **k-sets**: all subsets of $\{1, \ldots, n\}$ of cardinality k. We denote this class by $\mathcal{C}_k$.

For any $t = 1, 2, \ldots$ denote by $\mathbb{P}_\emptyset$ the distribution of U^t under the null, and by $\mathbb{P}_S$ the distribution under the alternative with contaminated set $S \in \mathcal{C}$. In addition, for a positive integer N, we denote by $\mathbb{P}^{\otimes N}$ the product measure $\mathbb{P} \otimes \ldots \otimes \mathbb{P}$ with N factors, and we let as previously $[N] = \{1, \ldots, N\}$.

5.1.2 Adaptive vs. Non-adaptive Sensing and Testing

Clearly, the above hypothesis testing problem would be trivial if one has access to an infinite number of i.i.d. samples $(U^t)_{t \in \{1,\ldots,\infty\}}$. Therefore, one must include some further restrictions on the data that is made available for testing. In particular, we will only consider testing procedures that make use of at most M entries of the matrix $(U_i^t)_{t \in \{1,\ldots,\infty\}, i \in [n]}$. It is useful to regard this as a matrix with n columns and an infinite number of rows.

The key idea of adaptive sensing is that information gleaned from previous observations can be used to guide the collection of future observations. To formalize this idea consider the following notational choices: for any subset $A \subseteq [n]$, we write $|A| = \mathrm{card}(A)$ the cardinality of A. When A is nonempty we write $U_A = (U_i)_{i \in A} \in \mathbb{R}^{|A|}$ for the subvector of a vector $U \in \mathbb{R}^n$ indexed by coordinates in A. Finally, if U is a random variable taking values in $\mathbb{R}^n$ denote by $\mathbb{P}|_A$ the distribution of U_A.

Let $S \in \mathcal{C} \cup \{\emptyset\}$ be the set of contaminated coordinates, and $M \geq 2$ be an integer. In our model we are allowed to collect information as follows. We consider successive rounds. At round $t \in \mathbb{N}$, one chooses a non-empty *query* subset $A^t \subseteq [n]$ of the components, and observes $U^t_{A^t}$. To avoid technical difficulties later on, we define the observation made at time t as X^t, so that $X^t_{A^t} = U_{A^t}$ and $X^t_{[n] \setminus A^t} = \mathbf{0}$. In words, one observes the A^t coordinates of U^t, while the remaining coordinates are completely uninformative. Each successive round proceeds in the same fashion, under the requirement that the budget constraint

$$\sum_{t=1}^{\infty} |A^t| \leq M \tag{5.1}$$

is satisfied. Note that clearly, the number rounds is no larger than M. Again, to avoid technical difficulties we assume the total number of rounds to be M in what follows, even if this means $A^t = \emptyset$ for some values of t. Figure 5.1 illustrates how information can be obtained within the sensing model for $n = 60$, under the alternative hypothesis with $S = [10]$ (i.e., the leftmost 10 components are contaminated). The query sequence is $A^1 = [60]$, $A^2 = [30]$, $A^3 = \{6, \ldots, 15\} \cup \{25, \ldots, 30\}$, corresponding to a budget M of at least $|A^1| + |A^2| + |A^3| = 106$ coordinate measurements.

In our setting, one can select the query sequence randomly and sequentially, and hence, we write the query sequence $(a^1, \ldots, a^M)$ as a realization of a sequence $(A^1, \ldots, A^M)$ of M random subsets of $[n]$, some of which may be empty, and such that $\sum_{t=1}^{M} |A^t| \leq M$.

A key aspect of adaptive sensing is that the query at round T may depend on all the information available up to that point. We assume A^t can depend on the history at time

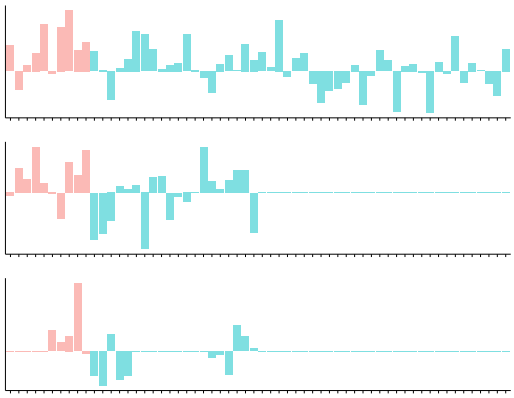

Figure 5.1: Illustration of a query sequence: samples from X^1 (top), X^2 (middle), X^3 (bottom)

$t-1$, which we denote by $H^{t-1} = (A^j, X^j)_{j\in[t-1]}$. More precisely, we assume A^t is a measurable function of H^{t-1}, and possibly of additional randomization. We call the collection of all the conditional distributions of A^t given H^{t-1} for $t \in [M]$ the *sensing strategy*. In particular, if there is no additional randomization, A^t is a deterministic function of H^{t-1}. We denote the set of all possible adaptive sensing strategies with sensing budget M as $\mathbf{AS}(M)$.

At this point it important to formally also clarify what is meant by *non-adaptive sensing*. This is simply the scenario where $(A^t)_{t\in[M]}$ is independent of $(U_i^t)_{t\in[M]}, i \in [n]$. In other words, all the decisions regarding the collection of data must be taken before any observations are made. The collection $(A^t)_{t\in[M]}$ is known as a *non-adaptive sensing strategy*, and the collection all such strategies satisfying the sensing budget (5.1) is denoted by $\mathbf{NAS}(M)$. A natural and important choice is *uniform sensing*, where $A^t = [n]$ for $t = 1,\dots,M/n$ (assume M is divisible by n). In words, one collects $m = M/n$ i.i.d. samples from $\mathbb{P}_S$. This problem has been thoroughly studied in (Arias-Castro et al., 2014); we summarize some of the main results of (Arias-Castro et al., 2014) in Section 5.1.3.

Now that we have formalized how data is collected, we can perform statistical tests. Formally, a *test* is a measurable binary function $\phi : H^M \mapsto \phi(H^M) \in \{0,1\}$, that is, a binary function of all the information obtained by the (adaptive or non-adaptive) sensing strategy. The result of the test is $\phi(H^M)$, and if this is one we declare the rejection of the null hypothesis. Finally, an *adaptive testing procedure* is a pair $(\mathcal{A},\phi)$ where $\mathcal{A}$ is a sensing strategy and ϕ is a test.

For any sensing strategy $\mathcal{A}$ and $S \in \mathcal{C}$, define $\mathbb{P}_\emptyset^{\mathcal{A}}$ (resp. $\mathbb{P}_S^{\mathcal{A}}$) as the distribution under the null (resp. under the alternative with contaminated set S) of the joint sequence $(A^1, X^1, \dots, A^M, X^M)$ of queries and observations. The performance of an adaptive testing procedure $(\mathcal{A},\phi)$ is evaluated by comparing the worst-case risk

$$R(\mathcal{A},\phi) = \mathbb{P}_\emptyset^{\mathcal{A}}(\phi \neq 0) + \max_{S\in\mathcal{C}} \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1)$$

to the corresponding minimax risk $R^*_{\mathbf{AS}} = \inf_{\mathcal{A}\in\mathbf{AS}(M),\phi} R(\mathcal{A},\phi)$, where the infimum is over all adaptive testing procedures $(\mathcal{A},\phi)$ with a budget of M coordinate measurements. The minimax risk $R^*_{\mathbf{AS}}$ depends on M, although we do not write this dependence explicitly

for notational ease. Likewise, the non-adaptive minimax risk $R^*_{\mathbf{NAS}}$ can be defined in an analogous way.

Let $m = M/n$ be the equivalent number of *full vector measurements*. In the following, we will just say m *measurements* for simplicity. This change of parameters allows for easier comparison with the special case of uniform sensing, where a full vector of length n is measured m times. In particular, when $m = M/n$ is an integer, uniform sensing corresponds to the deterministic sensing procedure with $A^t = [n]$ for $t \in [m]$, $A^t = \emptyset$ for $t > m$, and $\mathbb{P}_S^A = \mathbb{P}_S^{\otimes m}$ for $S \in \mathcal{C} \cup \{\emptyset\}$.

We are interested in the *high-dimensional* setting, where the ambient dimension n is high. All quantities such as the correlation coefficient ρ, the contaminated set size k, and the number of vector measurements m will thus be allowed to depend on n. In particular, we always assume that n, k and m all go to infinity simultaneously, albeit possibly at different rates, and our main concern is to identify the range of parameters in which it is possible to construct adaptive tests whose risks converge to zero. We consider the sparse regime where $k = o(n)$. Although the case of fixed ρ is of interest, most of our results will be concerned with the case where ρ converges to zero with n. When $\rho = 1$, the problem is trivial as detecting duplicate entries in a single sample vector from the distribution allows one to perform detection perfectly, while for fixed $\rho < 1$, the problem essentially becomes easier as the measurement budget m increases.

5.1.3 Uniform Sensing and Testing

The simplest and most-natural type of non-adaptive sensing strategy we can consider is uniform sensing. As stated before, this corresponds to the choice $A^t = [n]$ for $t = 1, \dots, m$ (recall that $m = M/n$), that is one collects m i.i.d. samples from $\mathbb{P}_S$. The minimax risk and the performance of several uniform sensing testing procedures have been analyzed in (Arias-Castro et al., 2014). The authors of that work analyzed the performance of tests based on the *localized squared sum* statistic

$$T_{\mathrm{loc}} = \max_{S \in \mathcal{C}} \sum_{t=1}^{m} \left(\sum_{i \in S} X_i^t \right)^2,$$

which was shown to be near-optimal in a variety of scenarios. The localized squared sum test that rejects the null hypothesis when T_{loc} exceeds a properly chosen threshold was shown to have an asymptotically vanishing risk when, for some positive constant c,

$$\rho k \geq c \max\left(\sqrt{\frac{\log |\mathcal{C}|}{m}}, \frac{\log |\mathcal{C}|}{m} \right). \tag{5.2}$$

This condition was shown to be near-optimal in most regimes for the classes of k-sets and k-intervals, unless k exceeds $\sqrt{n}$. In this latter and rather easier case, the simple non-localized squared sum statistic $T_s = \sum_{t=1}^{m} \left(\sum_{i=1}^{n} X_i^t \right)^2$ is near optimal. From (5.2), it is easy to see that the size of the class plays an important role, as a smaller class $\mathcal{C}$ leads to a weaker sufficient condition for detection. In particular, the localized squared sum test has asymptotically vanishing risk when

$$\textbf{k-sets: } \rho \geq c \max\left(\sqrt{\frac{\log n}{km}}, \frac{\log n}{m} \right), \quad \textbf{k-intervals: } \rho \geq c \max\left(\frac{1}{k}\sqrt{\frac{\log n}{m}}, \frac{\log n}{km} \right).$$

Necessary conditions for detection almost matching the previous sufficient conditions have been derived in (Arias-Castro et al., 2014). Although the dependence on the ambient dimension n is only logarithmic, this can still be significant in regimes where n is large but m is small.

5.1.4 Related Work

A closely related problem is that of detecting non zero mean components of a Gaussian vector X, referred to as the *detection-of-means* problem. This problem has received ample attention in the literature, see, for instance, (Ingster, 1997; Baraud, 2002; Donoho and Jin, 2004; Arias-Castro et al., 2008; Addario-Berry et al., 2010; Hall and Jin, 2010) and references therein. The detection-of-means problem can be formulated as the multiple hypothesis testing problem

$$\begin{aligned} H_0: &\quad X \sim \mathcal{N}(0, I_n), \\ H_1: &\quad X \sim \mathcal{N}(\mu \mathbf{1}_S, I_n), \text{ for some } S \in \mathcal{C}\ . \end{aligned}$$

where $\mathbf{1}_S$ is the indicator vector of S, I_n is the identity matrix, and $\mu \neq 0$. In other words, one needs to decide whether the components of X are independent standard normal random variables or they are independent normals with unit variance, and there is a (unknown) subset S of k components that have non-zero mean. The set of contaminated components S is assumed to belong to a class $\mathcal{C}$ of subsets of $[n]$. The behavior of the minimax risk has been analyzed for various class choices $\mathcal{C}$ (Ingster, 1997; Butucea et al., 2013; Arias-Castro et al., 2008; Addario-Berry et al., 2010). Detection and estimation in this model has been analyzed under adaptive sensing in (Castro, 2012; Haupt et al., 2009), where it is shown that, perhaps surprisingly, all sufficiently symmetric classes $\mathcal{C}$ lead to the same almost matching necessary and sufficient conditions for detection. This is quite different from the non-adaptive version of the problem where size and structure of $\mathcal{C}$ influence, in a significant way, possibilities of detection (see (Addario-Berry et al., 2010)).

Observe that the correlation model of Section 5.1.1 can be rewritten as

$$\begin{aligned} H_0: &\quad U_i^t = Y_i^t, i \in \{1, \dots, n\}, \\ H_1: &\quad U_i^t = \begin{cases} Y_i^t, & i \notin S, \\ \sqrt{1-\rho} Y_i^t + \sqrt{\rho} N^t, & i \in S \end{cases} \text{ for some } S \in \mathcal{C}\ . \end{aligned}$$

with $(Y_i^t), N^t$ independent standard normals, and that, as a consequence, the correlation model can be seen as a *random mean shift* model, with a slightly different normalization. However, most results on adaptive sensing for detection-of-means heavily hinge on the independence assumption between coordinates, which is not applicable for the detection of correlations. In particular, we shall see that the picture is more subtle in the presence of correlations.

A second problem, perhaps even more related, is that of detection in sparse principal component analysis (sparse PCA) within the *rank one spiked covariance model*, defined as the testing problem

$$\begin{aligned} H_0: &\quad X \sim \mathcal{N}(0, I_n), \\ H_1: &\quad X \sim \mathcal{N}(0, I_n + \theta u u^T), \text{ for some } u \in \mathbb{R}^n \text{ with } \|u\|_0 = k\ , \|u\|_2 = 1\ , \end{aligned}$$

where $\|u\|_0$ is the number of nonzero elements of u, and $\|u\|_2$ is the Euclidean norm of u. There is, also for this problem, a growing literature, see (Johnstone and Lu, 2009; Berthet and Rigollet, 2013; Cai et al., 2013). Note that when the coordinates of u are constrained in $\{0, 1/\sqrt{k}\}$, we recover a problem akin to that of detection of positive correlations, but with *unnormalized variances* over the contaminated set. The related problem of support estimation has been considered in (Amini and Wainwright, 2008) under the similar assumption that coordinates of u are constrained in $\left\{0, \pm 1/\sqrt{k}\right\}$.

5.1.5 Outline

The main contribution of this work is to show that adaptive sensing procedures can significantly outperform the best non-adaptive tests for the model in Section 5.1.1. We tackle the classes of k-intervals and k-sets. For k-intervals, necessary and sufficient conditions are almost matching. In particular, the number of measurements m necessary and sufficient to ensure that the risk approaches zero has essentially no dependence on the signal dimension n. This is in stark contrast with the non-adaptive sensing results, where it is necessary for m to grow logarithmically with n.

For k-sets, we obtain sufficient conditions that still depend logarithmically in n, but which improve nonetheless upon uniform sensing in some regimes. Although not uniform, the proposed sensing strategy is still non-adaptive. In addition to this, in a slightly different model akin to that of sparse PCA mentioned above, we show that all previous results (both non-adaptive and adaptive) carry on, and we obtain a tighter sufficient condition for detection of k-sets, that is nearly independent of the dimension n, and also improves significantly over non-adaptive sensing. Our results are summarized in Table 5.1. The chapter is structured as follows. We obtain a general lower bound in Section 5.2, and study various classes of contaminated sets. In Section 5.3, we propose procedures for k-sets and k-intervals. In Section 5.4, we prove a tighter sufficient condition under a slightly different model, for k-sets. Finally, we conclude with a discussion in Section 5.5. This chapter is joint work with Rui Castro and Gábor Lugosi (Castro et al., 2013).

		reference	$\rho k \to 0$	$\rho k \to \infty$
k-sets	necessary cdt.	Thm. 2	$\rho k\sqrt{m} \to \infty$	-
	sufficient cdt.	Prop. 20	$\rho\sqrt{km} \geq \sqrt{\log \frac{n}{k}}$, and $\rho km \geq \log \frac{n}{k}$	same
	suff. cdt. (unnormalized model)	Prop. 22	$\rho\sqrt{km} \geq \log\log \frac{n}{k}$	same
	suff. cdt. (uniform, $k = o(\sqrt{n})$)	(Arias-Castro et al., 2012)	$\rho\sqrt{km} \geq \sqrt{\log n}$, and $\rho m \geq \log n$	same
	necessary cdt. (uniform)	(Arias-Castro et al., 2012)	$\rho\sqrt{km} \geq \sqrt{\log \frac{n}{k^2}}$, and $\rho m \geq \log \frac{n}{k^2}$	same
k-int.	necessary cdt.	Thm. 2	$\rho k\sqrt{m} \to \infty$	-
	sufficient cdt.	Prop. 19	$\rho k\sqrt{m} \geq \sqrt{\log\log \frac{n}{k}}$	$\rho km \geq \log\log \frac{n}{k}$
	sufficient cdt. (uniform)	(Arias-Castro et al., 2012)	$\rho k\sqrt{m} \geq \sqrt{\log \frac{n}{k}}$	$\rho km \geq \log \frac{n}{k}$
	necessary cdt. (uniform)	(Arias-Castro et al., 2012)	$\rho k\sqrt{m} \geq \sqrt{\log \frac{n}{k}}$,	$\rho km \geq \log \frac{n}{k}$

Table 5.1: Summary of results (constants omitted).

5.1.6 Notation

We denote by $\mathbb{E}_{\mathbb{P}}$ the expectation with respect to a distribution $\mathbb{P}$. The Kullback-Leibler (KL) divergence between two probability distributions $\mathbb{P}$ and $\mathbb{Q}$ such that $\mathbb{P}$ is absolutely

continuous with respect to $\mathbb{Q}$ is $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \mathbb{E}_{\mathbb{P}}[\log(d\mathbb{P}/d\mathbb{Q})]$, with $d\mathbb{P}/d\mathbb{Q}$ the Radon-Nikodym derivative of $\mathbb{P}$ with respect to $\mathbb{Q}$. When $\mathbb{P}$ and $\mathbb{Q}$ admit densities f and g, respectively, with respect to the same dominating measure, we write $\mathrm{KL}(\mathbb{P}\|\mathbb{Q}) = \mathrm{KL}(f\|g)$. We denote by $\mathbf{1}_A$ the indicator function of an event or condition A.

5.2 Lower Bounds

We say that a sequence $z = (a^1, x^1, \ldots, a^M, x^M) \in \left(2^{[n]} \times \mathbb{R}^n\right)^M$ is *M-admissible* if $\sum_{t=1}^M |a^t| \leq M$. Consider an adaptive testing procedure $(\mathcal{A}, \phi)$, with query sequence $(A^1, \ldots, A^M) \in \left(2^{[n]}\right)^M$, and $(X^1, \ldots, X^M) \in (\mathbb{R}^n)^M$ the corresponding sequence of observations. Let $S \in \mathcal{C} \cup \{\emptyset\}$ be the set of contaminated coordinates. For $t \in [M]$, we denote by $f_{A^t|H^t}(\cdot|h^t)$ the probability mass function of A^t given $H^t = h^t$, and by $f_{X^t|A^t;S}(\cdot|a^t)$ the density of $X^t|A^t = a^t$ over $\mathbb{R}^n$ with respect to a suitable dominating measure over $\mathbb{R}^n$ (e.g., the product of Lebesgue measure and a point mass at 0). Therefore, the joint sequence $Z = (A^1, X^1, \ldots, A^M, X^M)$ admits a density f_S with respect to some appropriate dominating measure. For any M-admissible sequence $(a^1, x^1, \ldots, a^M, x^M)$, this density factorizes as

$$f_S(a^1, x^1, \ldots, a^M, x^M) = \prod_{t=1}^M f_{A^t|H^t}(a^t|a^1, x^1, \ldots, a^{t-1}, x^{t-1})\, f_{X^t|A^t;S}(x^t|a^t)\,.$$

For concreteness, let the density f_S be zero on any joint subsequence that is not M-admissible. It is crucial to note that all the terms in the factorization corresponding the sensing strategy do not depend on S. This is central to our arguments, as likelihood ratios simplify. More precisely, for any M-admissible sequence $(a^1, x^1, \ldots, a^M, x^M)$,

$$\frac{f_\emptyset(a^1, x^1, \ldots, a^M, x^M)}{f_S(a^1, x^1, \ldots, a^M, x^M)} = \prod_{t=1}^M \frac{f_{X^t|A^t;\emptyset}(x^t|a^t)}{f_{X^t|A^t;S}(x^t|a^t)} = \prod_{t=1}^M \frac{f_{X^t_{A^t}|A^t;\emptyset}(x^t_{a^t}|a^t)}{f_{X^t_{A^t}|A^t;S}(x^t_{a^t}|a^t)},$$

where the second equality follows from the sensing model.

Likelihood ratios play a crucial role in the characterization of testing performance. In particular, a classical argument (see, e.g., (Tsybakov, 2009, Lemma 2.6)) shows that, for any distributions $\mathbb{P}, \mathbb{Q}$ over a common measurable space Ω and any measurable function $\phi : \Omega \to \{0, 1\}$,

$$\mathbb{P}(\phi \neq 0) + \mathbb{Q}(\phi \neq 1) \geq \frac{1}{4}\exp\left(-\mathrm{KL}(\mathbb{P}\|\mathbb{Q})\right)\,.$$

Therefore

$$\begin{aligned} R^* = \inf_{(\mathcal{A},\phi)}\left[\mathbb{P}_0^{\mathcal{A}}(\phi \neq 0) + \max_{S\in\mathcal{C}} \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1)\right] &= \inf_{(\mathcal{A},\phi)} \max_{S\in\mathcal{C}}\left[\mathbb{P}_0^{\mathcal{A}}(\phi \neq 0) + \mathbb{P}_S^{\mathcal{A}}(\phi \neq 1)\right] \\ &\geq \inf_{(\mathcal{A},\phi)} \max_{S\in\mathcal{C}}\left[\frac{1}{4}\exp(-\mathrm{KL}(\mathbb{P}_0^{\mathcal{A}}\|\mathbb{P}_S^{\mathcal{A}}))\right] \\ &= \frac{1}{4}\exp(-\sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}}\|\mathbb{P}_S^{\mathcal{A}})). \end{aligned}$$

This entails that the minimax risk under adaptive sensing can be lower bounded by upper bounding the maximin KL divergence. Here, in order to bound the maximum KL divergence, we will take an approach similar to (Castro, 2012) for detection-of-means under

adaptive sensing, although our setup differs slightly. In (Castro, 2012), the testing procedures measure a single coordinate at a time, while we need multiple measures per step in order to capture correlations. We have the following necessary condition.

Theorem 2. *Let $\mathcal{C}$ be either the class of k-sets or k-intervals or disjoint k-intervals, and define*

$$D(\rho,k) = \min\left[\frac{\rho}{2(1-\rho)}, \rho^2(k+1)\right].$$

Then the minimax risk $R^*_{\boldsymbol{AS}}$ of adaptive testing procedures with a measurement budget of $M = mn$ coordinates is lower bounded as

$$R^*_{\boldsymbol{AS}} \geq \frac{\exp\left(-mkD(\rho,k)\right)}{4}.$$

Proof. First remark the following: for $\rho \leq 1/2$, and for any $A \subseteq [n]$,

$$\mathrm{KL}(\mathbb{P}_0|_A \,\|\, \mathbb{P}_S|_A) \leq D(\rho,k)\,|A \cap S|\,.$$

The proof is given in Appendix 5.6.2. The KL divergence between the joint probability models can we written as

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \,|\, \mathbb{P}_S^{\mathcal{A}}) &= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\log \frac{f_{X^t_{A^t}|A^t;\emptyset}(x^t_{A^t}|A^t)}{f_{X^t_{A^t}|A_t;S}(x^t_{A^t}|A^t)}\middle| A^t\right]\right] \\
&= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\mathrm{KL}(f_{X^t_{A^t}|A^t;\emptyset}(\cdot|A^t) \,\|\, f_{X^t_{A^t}|A^t;S}(\cdot|A^t))\right] \\
&= \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[\mathrm{KL}(\mathbb{P}_0|_{A^t} \,\|\, \mathbb{P}_S|_{A^t})\right] \\
&\leq D(\rho,k) \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}\left[|A^t \cap S|\right] \\
&= D(\rho,k) \sum_{i\in S} b_i
\end{aligned}$$

using the shorthand $b_i = \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}[\mathbf{1}_{i\in A^t}]$. Hence,

$$\sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \,|\, \mathbb{P}_S^{\mathcal{A}}) \leq D(\rho,k) \sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \sum_{i\in S} b_i\,.$$

Define the *class complexity*

$$\mathfrak{C}(\mathcal{C},M) = \sup\left\{\min_{S\in\mathcal{C}} \sum_{i\in S} b_i \,:\, b\in\mathbb{R}^n_+, \sum_{i=1}^{n} b_i \leq M\right\}.$$

For any sensing strategy $\mathcal{A}$, it holds that $\sum_{i=1}^{n} b_i = \sum_{t=1}^{M} \mathbb{E}_{\mathbb{P}_0^{\mathcal{A}}}[|A^t \cap S|] \leq M$, such that

$$\sup_{\mathcal{A}} \min_{S\in\mathcal{C}} \mathrm{KL}(\mathbb{P}_0^{\mathcal{A}} \,|\, \mathbb{P}_S^{\mathcal{A}}) \leq D(\rho,k)\,\mathfrak{C}(\mathcal{C},M)\,.$$

From (Castro, 2012, Lemma 3.1), we conclude that, for the both classes $\mathcal{C}_k$ and $\mathcal{D}_{[k]}$, respectively k-sets and disjoint k-intervals we have $\mathfrak{C}(\mathcal{C}_k, M) = \mathfrak{C}(\mathcal{D}_{[k]}, M) = \frac{Mk}{n} = mk$ (assuming without loss of generality that n/k is an integer[1]). As $\mathfrak{C}(\cdot, M)$ is decreasing with respect to set inclusion for any fixed M, $\mathfrak{C}(\mathcal{C}_{[k]}, M) = mk$ as well, and the result follows.

□

The lower bound argument in Theorem 2 yields the same lower bound for detection using any of the three classes of interest. This phenomenon is akin to what was observed in the context of detection-of-means under adaptive sensing, where the lower bounds are the same provided the classes of contaminated components are symmetric. In this setting, it was shown in addition in (Castro, 2012) that the condition in the lower bound is essentially sufficient and therefore, unlike in the non-adaptive counterpart of the problem, knowledge of the structure of $\mathcal{C}$ does not make the detection problem any easier. However, the problem of detection of correlations considered here seems to be more subtle in that one lacks matching upper bounds for all cases. Namely, we do not know whether: (a) for detection-of-correlations structure does not help; or (b) the lower bound is loose for some classes, in particular the class of k-sets.

Recall that we are interested in the characterization of the regimes for which the risk $R^*_{\mathbf{AS}}$ converges to zero as $m, k, n \to \infty$. Clearly, if ρ decays at a rate no faster than $1/k$, the previous necessary condition for the risk to vanish asymptotically is always satisfied. Nevertheless, the lower bound gives an indication about the rate at which the risk converges to zero. However, when $\rho = o(1/k)$ the situation is different, and Theorem 2 leads to the following necessary condition.

Corollary 1. *Let $\mathcal{C}$ denote either the class of k-sets, k-intervals or disjoint k-intervals, and suppose $\rho = o(1/k)$. For $R^*_{\mathbf{AS}}$ to converge to zero it is necessary that $\rho k\sqrt{m} \to \infty$.*

Proof. From the previous results, it is necessary that

$$mk \min\left[\frac{\rho}{2(1-\rho)}, \rho^2(k+1)\right]$$

goes to infinity for the risk to converge to zero. This quantity is bounded by $m\rho^2k^2$ asymptotically, and $m\rho^2k^2 \to \infty$ if and only if $\rho k\sqrt{m} \to \infty$. □

Recall that a sufficient condition for non-adaptive detection of k-intervals with the localized squared sum test is

$$\rho k\sqrt{m} > c\sqrt{\log(n)} \text{ and } \rho km > c\log(n)\ .$$

When $\rho = o(1/k)$ one has, asymptotically, $\rho k < 1$ and the first condition is stronger than the second. Non-adaptive detection with k-intervals is thus possible asymptotically for $\rho k\sqrt{m} > c\sqrt{\log(n)}$. This corresponds to the condition of Corollary 1 up to a logarithmic factor in n, which implies that in the case of k-intervals, one can improve at most by a factor logarithmic in n with adaptive sensing. This can be still quite significant, however, and we show in Section 5.3 that this can indeed be achieved.

[1] If n/k is not an integer, one can directly show that $\mathfrak{C}(\mathcal{D}_{[k]}, M) \leq 2mk$ and the result of the theorem follows with mk replaced by $2mk$.

5.3 Adaptive Tests

5.3.1 Sequential Thresholding

In the context of support recovery from signals with independent entries using adaptive sensing, (Malloy and Nowak, 2011b,a) have proposed the sequential thresholding (ST) procedure, which is based on an intuitive bisection idea. Although initially introduced for support estimation, ST can be easily adapted to detection, and we present such results here. In addition, we present a slight generalization to signals with independent *vector entries*.

Let $\mathbb{Q}_0$ and $\mathbb{Q}_1$ be two probability distributions over $\mathbb{R}^{\tilde{d}}$, and let $Z \in \mathbb{R}^{\tilde{n}\times\tilde{d}}$ be a random matrix. Consider the multiple testing problem defined as follows. Under the null, Z has rows identically distributed according to $\mathbb{Q}_0$ Under the alternative, a small unknown subset of $\tilde{k}$ rows of Z are distributed according to $\mathbb{Q}_1$, while the remaining rows are distributed according to $\mathbb{Q}_0$. In both cases, all rows are independent. More formally, denote by $Z_1, \dots, Z_{\tilde{n}}$ the rows of Z, such that the testing problem is

$$
\begin{aligned}
&H_0 : Z_i \sim \mathbb{Q}_0^{\otimes \tilde{n}}, \\
&H_1 : Z_i \sim \mathbb{Q}_0 \text{ for } i \notin S, \quad Z_i \sim \mathbb{Q}_1 \text{ for } i \in S, \quad \text{for some } S \in \mathcal{C} \text{ with } |S| = \tilde{k},
\end{aligned}
$$

where, as already mentioned, all rows are independent in both cases. We refer to this testing problem as that of detection from *signals with independent (vector) entries*. The framework of adaptive sensing introduced in Section 5.1.2 can be easily adapted to this model. In this case, in order to allow for vector entries, we consider that the experimenter is allowed to obtain samples from rows of Z, and that he can select which rows to query in a sequential manner as previously, under the constraint that the total number of rows measured be less than M. We also refer to this straightforward extension as adaptive sensing, and we say that $\tilde{m} = M/\tilde{n}$ is the number of *measurements* (i.e., $\tilde{m}$ is the equivalent number of times the full matrix Z was observed).

Sequential thresholding is a procedure for testing with adaptive sensing within the type of model just mentioned. Assume that $\mathbb{Q}_0$ and $\mathbb{Q}_1$ admit densities f_0 and f_1, respectively, with respect to some common dominating measure, and for $i \in [n]$, denote by

$$
LR(f_1|f_0; z_i^1, \dots, z_i^{\tilde{m}}) = \frac{\prod_{t=1}^{\tilde{m}} f_0(z_i^t)}{\prod_{t=1}^{\tilde{m}} f_1(z_i^t)}
$$

the likelihood ratio associated to i.i.d. observations $z_i^1, \dots, z_i^{\tilde{m}} \in \mathbb{R}^{\tilde{d}}$ of Z_i, the i-th row of Z. ST proceeds as outlined in Figure 5.3. Initially, ST measures all $\tilde{n}$ rows $\tilde{m}$ times, and throws away a fraction (of about half under the null) of the $\tilde{n}$ rows based on the values of the likelihood ratios. This is repeated with the remaining rows a number of times logarithmic in $\tilde{n}$, at which point ST calls detection if some coordinates have not been thrown away. This is illustrated in Figure 5.2.

The following result is easily deduced from the analysis of ST for support estimation.

Proposition 17 (Sufficient condition for ST). *Assume* $\tilde{k}/\tilde{n} \to 0$, *and*

$$
\liminf_{\tilde{n}\to\infty} \frac{\tilde{m}\,\mathrm{KL}(f_0 \| f_1)}{\log\log_2 \tilde{n}} > 1,
$$

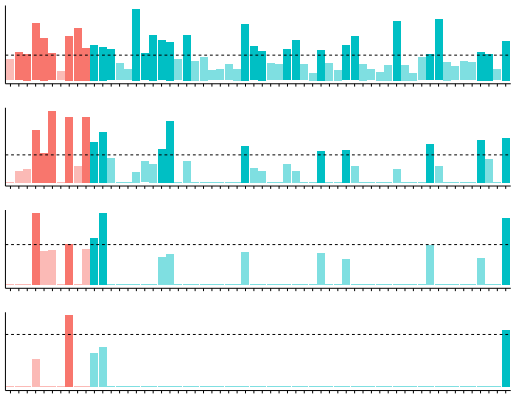

Figure 5.2: Illustration of sequential thresholding with $k = 10, n = 60$: contaminated coordinates are the first ten on the left. Bars depict likelihood ratios associated with each coordinate: at each step, coordinates with likelihood ratio below a threshold are thrown away.

Input: $K \simeq \log_2(\tilde{n})$ (number of steps),
$\gamma = \text{median}_{z_1^1,\dots,z_1^{\tilde{m}} \sim f_0}(LR(f_1|f_0; z_1^1, \dots, z_1^{\tilde{m}}))$ (threshold)
Initialization: $\mathcal{S}_0 = \{1, \dots, \tilde{n}\}$
for all $r = 1, \dots, K$ **do**
  for all $i \in \mathcal{S}_{r-1}$ **do**
    measure $z_i^1, \dots, z_i^{\tilde{m}} \sim Z_i$
    compute $LR_i = LR(f_1|f_0; z_i^1, \dots, z_i^{\tilde{m}})$
  end for
  $\mathcal{S}_r = \{i \in \mathcal{S}_{r-1} : LR_i > \gamma\}$
  if $\sum_{r=0}^{K} |\mathcal{S}_r| > 4\tilde{n}$ **then**
    return no detection
  end if
end for
return detection if $\mathcal{S}_K \neq \emptyset$

Figure 5.3: Sequential thresholding procedure

then the sequential thresholding procedure with a budget of $4\tilde{m}$ measurements has risk tending to zero as $\tilde{n}$ goes to infinity.

Proof. We begin by showing that the event of termination upon $\sum_{r=0}^{K} |\mathcal{S}_r| > 4\tilde{n}$ has an asymptotically vanishing probability. Assume the alternative hypothesis with contaminated set S. Then, similarly as in (Castro, 2012, Proposition 4.1), using Bernstein's inequality for sums of truncated hypergeometric variables,

$$P\left(\sum_{r=0}^{K} |\mathcal{S}_r| > 4\tilde{n}\right) \leq \exp\left(-\frac{\tilde{n} - \tilde{k}}{4 + \frac{2K}{3}}\right),$$

which converges to zero. The application of Chernoff-Stein's lemma as in (Malloy and

Nowak, 2011a) allows us to bound the probability of error as follows. The type I error of the procedure is bounded by

$$\frac{\tilde{n}-\tilde{k}}{2^K}.$$

Let $E_{i,t}$ denote the event that the likelihood ratio is below γ for coordinate i at step t (in which case, coordinate i will not be included in $\mathcal{S}_t$). Without loss of generality, assume that $1 \in S$. The type II error is

$$Q_1\left(\cap_{i\in S}\left(\cup_{t=1}^{K} E_{i,t}\right)\right) \leq \left(KQ_1\left(E_{1,1}\right)\right)^{\tilde{k}}.$$

We write $a = e^{-\tilde{m}D}$ for $\lim_{\tilde{m}\to\infty} \frac{\log a}{\tilde{m}} = D$. From the Chernoff-Stein lemma,

$$Q_1\left(E_{1,1}\right) = e^{-\tilde{m}\,\mathrm{KL}(f_0\|f_1)}.$$

Hence, for $K = (1+\varepsilon_1)\log_2 n$ and $\varepsilon_2 > 0$, there exists $\tilde{m}_0$ such that for $\tilde{m} \geq \tilde{m}_0$, the type II error is bounded by

$$\left(Ke^{-\tilde{m}(\mathrm{KL}(f_0\|f_1)-\varepsilon_2)}\right)^{\tilde{k}} = \exp\left(\tilde{k}\log\left[(1+\varepsilon_1)\log_2 n\right] - \tilde{m}\tilde{k}(\mathrm{KL}(f_0\|f_1)-\varepsilon_2)\right).$$

Hence, the risk goes to zero if for some $\varepsilon_1, \varepsilon_2 > 0$, it holds that

$$\liminf_{\tilde{n}\to\infty} \frac{\tilde{m}(\mathrm{KL}(f_0\|f_1)-\varepsilon_2)}{\log\left[(1+\varepsilon_1)\log_2 n\right]} > 1.$$

As a consequence, for the risk to go to zero, it is sufficient that

$$\liminf_{\tilde{n}\to\infty} \frac{\tilde{m}\,\mathrm{KL}(f_0\|f_1)}{\log\log_2 n} > 1.$$

□

Note that the ST procedure does not require knowledge of $\tilde{k}$. ST can be applied to the case of k-intervals, as we demonstrate in the next section.

5.3.2 The Case of k-intervals

In this section, we look at the case of the class $\mathcal{C}_{[k]}$ of intervals of length k. It is sufficient to work with the class $\mathcal{D}_{[k]}$ of disjoint intervals for the following reason: assume that one has a procedure for detection of disjoint k-intervals. Then, for detection of general k-intervals, *this procedure can be applied as if the objective was detection of disjoint $k/2$-intervals*. Indeed, if S is any k-interval, there exist at most two sets in $\mathcal{D}_{[k/2]}$ that intersect S, and at least one of them, say S', has a full intersection with S, i.e., $|S \cap S'| = k/2$. As a consequence, under mild conditions on the procedure, this leads to a sufficient condition for detection of k-intervals identical up to constants to that associated with the original procedure for disjoint k-intervals. In the following, we show how to perform detection in the case of disjoint k-intervals.

Recall that $\mathcal{D}_{[k]} = \{I_1, \ldots, I_{\lfloor n/k \rfloor}\}$, where $I_j = \{(j-1)k+1, \ldots, jk\}$ for $j \in [\lfloor n/k \rfloor]$. For simplicity, we will assume that n/k is an integer. In order to apply ST, we will treat intervals as n/k independent k-dimensional observations. Define $\tilde{n} = n/k$, $\tilde{k} = 1$, $\tilde{m} = m$,

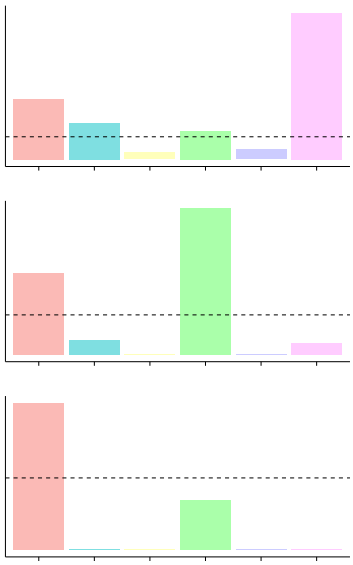

Figure 5.4: Illustration of sequential thresholding for k-intervals, with $n/k = 6$ intervals of size k. Bars depict likelihood ratios associated with the intervals.

and $\tilde{d} = k$. Let $\mathbb{Q}_0 = \mathbb{P}_0|_{I_1}$ be the joint probability distribution over an interval under the null, and $\mathbb{Q}_1 = \mathbb{P}_{I_1}|_{I_1}$ be the joint probability distribution over the contaminated interval under the alternative. We refer to the corresponding sequential thresholding procedure as *ST for disjoint k-intervals*. This procedure is illustrated in Figure 5.4. This provides the following sufficient condition for detection of disjoint k-intervals.

Proposition 18. *Assume that ρ converges to zero, and that either*

$$\rho k \to \infty \quad \text{and} \quad m \log(1+\rho k) \geq C_3 \log\log(n/k),$$

or

$$\rho k \to 0 \quad \text{and} \quad \rho k \sqrt{m} \geq C_4 \sqrt{\log\log(n/k)}$$

for some constants C_3 and C_4. Then sequential thresholding for disjoint k-intervals has risk converging to zero.

Proof. The detailed computations can be found in Appendix 5.6.3. Assume that $\rho k > 1$, then

$$\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) \geq \frac{\log(1+\rho k)}{10}.$$

Similarly, when $\rho k < 1/2$ and $k > 32$,

$$\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) \geq \frac{\rho^2 k^2}{16}.$$

Combined with Proposition 17, this gives the desired result. □

Consider the case where $\rho k \to \infty$. In that case, omitting constant factors, sequential thresholding would succeed for $m \geq \frac{\log\log(n)}{\log(1+\rho k)}$. Recall that uniform non-adaptive testing is possible for $m \geq \frac{c \log n}{\rho k}$. When $\rho k > \log(n)$ asymptotically, both conditions are trivially satisfied for m constant, while when $\rho k < \log(n)$, we already improve upon non-adaptive tests. In spite of this, the dependence on ρk of our sufficient condition when $\rho k \to \infty$ is logarithmic, while it is only linear for $\rho k \to 0$. This may appear surprising, as one may argue the former case corresponds to a regime where the signal is stronger (and so the problem should be easier). However, this surprising fact is solely an artifact from the sequential thresholding procedure, and can be fixed through a small modification of the sensing methodology.

In order to recover the same linear dependence in both cases, we propose to add a subsampling stage prior to sequential thresholding. This subsampling can be decided before any data is collected, and thus can be viewed as a non-adaptive aspect of the entire procedure. Consider the simple subsampling scheme wherein one keeps p coordinates per interval, for some $p \in \{2, \dots, k\}$, and measures each p-tuple $\left\lfloor \frac{mn}{pn/k} \right\rfloor = \left\lfloor \frac{mk}{p} \right\rfloor$ times. This prompts the following question: is there a value of p that allows one to detect more easily? Define the p-truncated intervals as $I_j^p = \{(j-1)k+1, \dots, (j-1)k+p\}$ for $j \in [n/k]$. Formally, we consider the deterministic sensing strategy $\mathcal{A}_p = (A^t)$ where for $t \in \left[\left\lfloor \frac{mk}{p} \right\rfloor\right]$,

$$A^t = \bigcup_{j \in [n/k]} I_j^p.$$

As this involves one simple testing problem per interval, the difficulty of testing is essentially characterized by the KL divergence $\mathrm{KL}(\mathbb{P}_0^{\mathcal{A}_p} \| \mathbb{P}_S^{\mathcal{A}_p})$ between the distributions under the null and the alternative. In this section, we make explicit the dependence of $\mathbb{P}_S$ and p by using the notation $\mathbb{P}_S^p$. Consider any fixed $S \in \mathcal{D}_{[k]}$, then the best KL divergence that can be obtained is

$$\max_{p \in \{2,\dots,k\}} \mathrm{KL}\left(\mathbb{P}_0^{\mathcal{A}_p} \| \mathbb{P}_S^{\mathcal{A}_p}\right) = \max_{p \in \{2,\dots,k\}} \left[\sum_{t=1}^{\left\lfloor \frac{mk}{p} \right\rfloor} \mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p) \right] = \left\lfloor \frac{mk}{p} \right\rfloor \max_{p \in \{2,\dots,k\}} \mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p),$$

which is independent of S . Due to nonlinearity in the KL divergence the optimal value of p is generally different than k, as illustrated in Figure 5.5. The optimal p and corresponding optimal value seem hard to compute analytically, but numerical evidence shows that, for ρ away from zero, the optimal p is of the order of ρ^{-1}. This observation is sufficient for our purposes, and is formalized below. Remark that when $\rho k < 1$, the optimal value of p is clamped to k.

Equipped with this subsampling stage when $\rho k \to \infty$, we can now modify the ST for k-intervals procedure as follows: when $\rho k \to \infty$, set $\tilde{m} = \left\lfloor \frac{mk}{p} \right\rfloor$, $\tilde{d} = \left\lceil \frac{1}{\rho} \right\rceil$, and use only observations corresponding to $\tilde{d}$ coordinates per interval. We refer to this new procedure as the *modified sequential thresholding for disjoint k-intervals*.

Proposition 19. *Assume that ρ converges to zero, and that either*

$$\rho k \to \infty \quad \text{and} \quad \rho k m \geq C_5 \log\log(n/k),$$

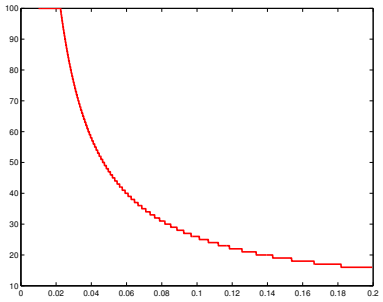

Figure 5.5: Optimal p as a function of ρ, for $k = 100$

or

$$\rho k \to 0 \quad \textit{and} \quad \rho k\sqrt{m} \geq C_6\sqrt{\log\log(n/k)}$$

for some constants C_5, C_6. Then the modified sequential thresholding for disjoint k-intervals has risk converging to zero.

Proof. We have the following straightforward new lower bound, proved in Appendix 5.6.4. Although the lower bound appears weaker than previously, this corresponds to a setting where more measurements can be carried out. When $\rho k > 1$, then, using subsampling with $p = \left\lceil \frac{1}{\rho} \right\rceil$,

$$\mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p) \geq \frac{1}{11}.$$

The sufficient condition for ST leads to the result. □

The adaptive procedure allows us to obtain a mild dependence on the original dimension n of the problem. When $\rho = o(1/k)$, this sufficient condition almost matches the lower bound of Corollary 1, while when $\rho k \to \infty$, the sufficient condition is already satisfied for $m = \log\log(n/k)$.

5.3.3 The Case of k-sets: Randomized Subsampling

In this section, we consider the class $\mathcal{C}_k$ of k-sets. In this case, we do not currently know whether a procedure along the lines of ST can be successfully applied. However, the idea of subsampling the coordinates can still be used to yield modest but important performance gains. While for disjoint k-intervals a deterministic subsampling was sufficient, this is not the case for k-sets, where any deterministic subsampling that selects less than about $n-k$ coordinates cannot have risk converging to zero. For this reason, we consider a *randomized* subsampling of the coordinates.

Consider a sample B of $\left\lfloor \frac{2np}{k} \right\rfloor$ elements drawn without replacement from $[n]$ for some $p \geq 2$. Let $\theta : \mathbb{R}^{\lfloor np/k \rfloor} \to \{0,1\}$ be the localized squared sum test with ambient dimension $\left\lfloor \frac{2np}{k} \right\rfloor$, and contaminated sets $\mathcal{C} = \mathcal{C}_{\lfloor p \rfloor}$ of size $\lfloor p \rfloor$, and consider the sensing strategy defined by

$$A^1 = \ldots = A^{\left\lfloor \frac{mk}{2p} \right\rfloor} = B.$$

We refer to the adaptive sensing procedure $((A^t), \theta)$ as the *randomized testing procedure*. Define $Y = |B \cap S|$ (resp. $Y = 0$) under the alternative with contaminated $S \in \mathcal{C}_k$ (resp. under the null), which is the number of contaminated elements in the subsample. Clearly Y is a hypergeometric random variable with expectation $\frac{k}{n}\left\lfloor \frac{2n}{k} p \right\rfloor \in [2p - k/n, 2p]$. In words, we consider a subsample of the coordinates, with about $2p$ contaminated coordinates (in expectation) under the alternative, and we apply the (non-adaptive) localized squared sum test.

Note that the procedure is strictly non-adaptive, as the subsampling can be decided in advance. However, this sensing strategy is a bit different than uniform sensing, as not all coordinates are measured. Nonetheless, this allows one to detect under weaker conditions than with uniform non-adaptive sensing when k is large enough.

Proposition 20. *Let $2 \le p \le k$ such that p goes to infinity. Assume that ρ converges to zero and that*

$$\rho mk \ge \frac{C_1}{\left[1 - \frac{1}{m} - \frac{1}{k}\right]} \log \frac{2pn}{k}, \quad \text{and} \quad \rho\sqrt{mk} \ge \frac{C_1}{\sqrt{(p-1)\left[1 - \frac{1}{m} - \frac{1}{k}\right]}} \sqrt{\log \frac{2pn}{k}},$$

for some constant C_1, then the randomized testing procedure has risk converging to zero.

Proof. Let η_I (resp. η_{II}) be the risk of type I (resp. of type II) for θ. The type I error of the randomized testing procedure is $p_I = \eta_I$. Let $p_+ = P(Y \ge \lfloor p \rfloor)$ the probability of the sample containing at least $\lfloor p \rfloor$ contaminated elements, and $p_- = 1 - p_+$. Note that since $\frac{2np}{k}\frac{k}{n} = 2p$ goes to infinity, we can assume that Y is distributed according to a Poisson distribution with parameter $2p$, as this is asymptotically equivalent to the hypergeometric distribution. Hence, we have $p_- = P(Y < \lfloor p \rfloor) \le \left(1 + \frac{p(2p)^p}{p!}\right)\exp(-2p)$. Using $p! \ge \sqrt{2\pi p}\left(\frac{p}{e}\right)^p$, we have that $p_- \le \exp(-2p) + \sqrt{p}\exp(-p/4)$, which converges to zero. The type II error of the randomized testing procedure is $p_{II} = p_+\eta_{II} + p_-(1 - \eta_I) \le \eta_{II} + p_-$. It remains to show that η_I and η_{II} both go to zero. This follows from the sufficient conditions for the localized squared sum test, and from $\lfloor p \rfloor \left\lfloor \frac{mk}{2p} \right\rfloor \ge \frac{mk}{2}\left[1 - 1/p + \frac{2(1-p)}{mk}\right] \ge \frac{mk}{2}\left[1 - 1/p - 1/m\right]$. Hence, the sufficient conditions for the localized squared sum test θ provides the result. □

In particular, for $p = \log\log n$, it is sufficient that, omitting constants,

$$\rho mk \ge \log \frac{n}{k}, \quad \rho\sqrt{mk} \ge \sqrt{\log \frac{n}{k}},$$

to ensure the detection risk converges to zero. This does not match the adaptive lower bound, and the dependence on n is still logarithmic. However, this already improves upon the setting of uniform non-adaptive sensing when $k \ge \frac{m}{\log n}$. Indeed, recall that using uniform sensing, the sufficient condition is

$$\rho\sqrt{mk} \ge \sqrt{\log n}, \quad \rho m \ge \log n.$$

The first condition is insensitive to subsampling, due to the dependence in mk, and we do not improve with respect to it. The second condition, however, only depends on m, and

does not get easier to satisfy when k is large. Hence, our result shows that it is more efficient when k is large enough to reduce to a problem with an almost constant contaminated set size, but with an increased budget of full vector measurements.

5.4 Unnormalized Correlation Model

5.4.1 Model and Extensions of Previous Results

An alternative choice to the previous correlation model is the following *unnormalized model* with covariance matrix

$$(\bar{\Sigma}_S)_{i,j} = \begin{cases} 1, & i = j, i \notin S, \\ 1+\rho, & i = j, i \in S, \\ \rho, & i \neq j, \text{and } i, j \in S, \\ 0 & \text{otherwise.} \end{cases}$$

under the alternative with contaminated set $S \in \mathcal{C}$. This model is a special case of the *rank one spiked covariance* model introduced in (Johnstone, 2001). Observe that this correlation model can also be rewritten as

$$\begin{aligned} H_0 : \quad & X_i^t = Y_i^t, i \in \{1, \dots, n\}, \\ H_1 : \quad & X_i^t = \begin{cases} Y_i^t, & i \notin S, \\ Y_i^t + \sqrt{\rho} N^t, & i \in S \end{cases} \text{for some } S \in \mathcal{C}. \end{aligned}$$

with $(Y_i^t), N^t$ independent standard normals. This can thus be interpreted as a random additive noise model, as for the model of Section 5.1.1. Observe that our original correlation detection model is obtained by normalizing each component such that the components have unit variance. This is a minor difference that does not essentially change the difficulty of detection in the non-adaptive setting (indeed all upper and lower bounds proved in (Arias-Castro et al., 2014) can be reproved for this model with minor modifications). Interestingly, however, under adaptive sensing the information provided by the higher variance in the contaminated components can be exploited to give a major improvement over the normalized model. This may be done by applying the sequential thresholding algorithm to the squares of the components as described below.

In the following, for any quantity X relative to the *normalized* model of Section 5.1.1, we denote by $\bar{X}$ the corresponding quantity related to the unnormalized model. All of previous results can be shown to hold for this model as well. As already mentioned, this includes the necessary and sufficient conditions of (Arias-Castro et al., 2014) (Proposition 26 in Appendix), but also the lower bound of Theorem 2 (Proposition 27 in Appendix), and sufficient conditions for k-sets and k-intervals of Propositions 20 and 19 (Proposition 29 in Appendix). In particular, the procedures associated to the sufficient conditions can be used with little modifications.

5.4.2 The Case of k-sets

The procedure proposed below combines randomized subsampling with sequential thresholding, in order to capitalize on the unnormalized model. Consider the second moments

$Y_i = X_i^2$. Under the alternative with contaminated set $S \in \mathcal{C}$, Y_i is distributed as follows: (a) for $i \notin S$, Y_i is distributed according to a chi-squared distribution with one degree of freedom (that we denote by χ_1^2), (b) for $i \in S$, Y_i is distributed as $(1+\rho)\,\chi_1^2$. Note that under our sensing model, it is perfectly legitimate to sample $A_1 = \{1\}, \dots, A_n = \{n\}$, and thus obtain independent samples of each of the coordinates of the random vector. In particular, this allows us to obtain independent samples from the coordinates of Y. As a consequence, we can directly apply ST to detect increased variance over a subset of the coordinates.

As already mentioned, ST does not require knowledge of k, which results in a sufficient condition that is independent of k. In particular, it does not become easier to satisfy as k increases. This condition can, however, be significantly weakened using the random subsampling used in last section. As in Proposition 20, this is due to the fact that by subsampling, one can increase the budget of full vector measurements, while the decrease in the contaminated set size does not impact the sufficient condition for detection. This is summarized in the following result, which can be proved similarly as Proposition 20.

Proposition 21 (Sufficient condition for ST+randomized subsampling)**.** *Assume* $\tilde{k}/\tilde{n} \to 0$, *and*

$$\liminf_{\tilde{n}\to\infty} \frac{\tilde{m}\tilde{k}\,\mathrm{KL}(f_0 \| f_1)}{(\log\log_2 \tilde{n})^2} > 1,$$

then the sequential thresholding procedure with randomized subsampling ($p = \log\log_2 \tilde{n}$) *and a budget of* $4\tilde{m}$ *full vector measurements has risk tending to zero as* $\tilde{n}$ *goes to infinity.*

Let $\tilde{n} = n$, $\tilde{k} = k$, and $\tilde{m} = m$. Let $\mathbb{Q}_0$ be the χ_1^2 distribution, and $\mathbb{Q}_1$ be the $(1+\rho)\,\chi_1^2$ distribution, both with respect to Lebesgue's measure. We consider the associated sequential thresholding procedure (with randomized subsampling), with the previous modification of sampling independent single coordinates. We refer to this procedure as *variance thresholding*. This leads to the following sufficient condition for detection.

Proposition 22. *Assume that* ρ *converges to zero and that*

$$\rho\sqrt{km} \geq C_2 \log\log_2 n$$

for some constant C_2. *Then, the risk of the variance thresholding procedure converges to zero.*

Proof. Let g be the density of a χ_1^2-distributed random variable, such that the density of a $(1+\rho)\chi_1^2$-distributed random variable is given by $\frac{1}{1+\rho} g\left(\frac{\cdot}{1+\rho}\right)$. Then, using $g(x) \propto x^{-1/2} e^{-x/2}$,

$$\begin{aligned}
\mathrm{KL}(\chi_1^2 \| (1+\rho)\chi_1^2) &= \int_{\mathbb{R}} \log\left(\frac{g(x)}{\frac{1}{1+\rho} g\left(\frac{x}{1+\rho}\right)}\right) g(x)\,dx \\
&= \log(1+\rho) + \int_{\mathbb{R}} \log\left(\frac{x^{-1/2} e^{-x/2}}{\left(\frac{x}{1+\rho}\right)^{-1/2} e^{\frac{-x}{2(1+\rho)}}}\right) g(x)\,dx \\
&= \log(1+\rho) + \int_{\mathbb{R}} \log\left(\frac{e^{\frac{-\rho x}{2(1+\rho)}}}{(1+\rho)^{1/2}}\right) g(x)\,dx \\
&= \frac{\log(1+\rho)}{2} - \frac{\rho}{2(1+\rho)} \int_{\mathbb{R}} x g(x)\,dx.
\end{aligned}$$

As the expectation of a χ_1^2-distributed random variable is one, this leads to

$$\mathrm{KL}(\chi_1^2 \| (1+\rho)\chi_1^2) = \frac{1}{2}\left[\log(1+\rho) - \frac{\rho}{1+\rho}\right] = \frac{\rho^2}{4} + o(\rho^2).$$

Plugging this expression into the sufficient condition of Proposition 21 provides the result. □

Assume for the following discussion that $\rho k \to 0$. The necessary condition that we have established previously is that $\rho k\sqrt{m}$ goes to infinity. Neglecting the double log factor, the sufficient condition that we have just obtained is that ρkm goes to infinity, which is stronger. Hence, there is a gap between the sufficient and necessary condition. In particular, that $\rho k\sqrt{m}$ goes to infinity was shown to be near-sufficient for detection with k-intervals, and the gap that we observe for k-sets does not allow us to conclude as to whether structure helps for detection (as is the case under non-adaptive sensing).

Recall that the unnormalized model is similar to that of detection in the problem of sparse PCA. The method of *diagonal thresholding* (also referred to as *Johnstone's diagonal method*) is a simple and tractable method for detection (and support estimation) in sparse PCA (with uniform non-adaptive sensing), which consists in testing based on the diagonal entries of empirical covariance matrix - that is, the empirical variances. Hence, it is similar to the method that we consider here, except that we estimate variances based on independent samples for each coordinate. Note that this last point is essential to our method. Indeed, consider the opposite case where we do not use independent samples for each coordinates. For the sake of illustration, assume $\rho = 1$, such that the contaminated components are exactly equal. In this case, the probability of throwing away one component is equal to that of throwing away *all* contaminated components, and failure will occur with fixed non small probability due to the use of dependent samples.

Finally, it is noteworthy that a naïve implementation of the optimal test in the non-adaptive setting has complexity $O(n^k)$, while with adaptive sensing, we obtain a procedure that can be carried out in time and space linear in n, and still improves significantly with respect to the non-adaptive setting.

5.5 Discussion

We showed that for k-intervals, adaptive sensing allows one to reduce the logarithmic dependence in n of sufficient conditions for non-adaptive detection to a mild $\log\log n$, and that this is near-optimal in a minimax sense.

For k-sets, the story is less complete. The sufficient condition obtained in the unnormalized model is still stronger than the sufficient condition obtained for k-intervals, and does not match our common lower bounds, which leaves open the question of *whether structure helps under adaptive sensing for detection of correlations*? The analogous question for detection-of-means has a negative answer, meaning structure does not provide additional information for detection. However, for detection-of-correlations a definite answer is still elusive. Another open question is to what extent adaptive sensing allows one to overcome the exponential computational complexity barrier that one can encounter in the non-adaptive setting.

Aside from the normalized and unnormalized correlation models, other types of models can be considered. A more general version of our normalized model has been analyzed

in (Arias-Castro et al., 2014), where the correlations need not be all the same, leading to results that involve the mean correlation coefficient $\rho_{\text{avg}} = \left(\sum_{i,j \in S : i \neq j} (\Sigma_S)_{i,j}\right) / k(k-1)$. In addition, we assume in most procedures that ρ and/or k are known, and it would be of interest to have procedures that do not require such knowledge.

5.6 Proofs

5.6.1 Inequalities and KL Divergences

In this section, we collect elementary inequalities that we use repeatedly in the computations.

$$\text{For } x > -1, \quad \log(1+x) \le x, \tag{5.3}$$

$$\text{For } x > 0, \quad \log(1+x) + \frac{1}{1+x} - 1 \le x^2, \tag{5.4}$$

$$\text{For } 0 < x < 1/2, \quad \log(1-x) + \frac{1}{1-x} - 1 \le 2x^2, \tag{5.5}$$

$$\text{For } x < 1, \quad -\log(1-x) - \frac{1}{1-x} + 1 \le x^2, \tag{5.6}$$

$$\text{For } x \in]-1,1], \quad \log(1+x) + \frac{1}{1+x} - 1 \ge \frac{x^2}{8}, \tag{5.7}$$

$$\text{For } x \ge 1, \quad \log(1+x) + \frac{1}{1+x} - 1 \ge \frac{\log(1+x)}{5}. \tag{5.8}$$

The following expression of the KL divergence is used throughout the chapter.

Proposition 23. *We have*

$$\begin{aligned} \mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_S) = \frac{\mathbf{1}_{k\ge 2}}{2}\Bigg[k\left(-1 + \frac{1}{1-\rho} + \log(1-\rho)\right) - \left(\frac{1}{1-\rho} + \log(1-\rho)\right) \\ + \left(\frac{1}{1+\rho(k-1)} + \log(1+\rho(k-1))\right)\Bigg]. \end{aligned} \tag{5.9}$$

Proof. The KL divergence between $\mathbb{P}_0$ and $\mathbb{P}_S$ can be computed using the standard formula for KL divergence between two centered Gaussian vectors, with covariance matrices

$$\Sigma_0 = I_n, \quad \Sigma_1 = \Sigma_S.$$

When $k < 2$, the divergence is zero, and we will thus assume $k \ge 2$. Up to a simultaneous permutation of rows and columns,

$$\Sigma_S = \begin{bmatrix} I_{n-k} & \\ & J_\rho(k) \end{bmatrix}$$

where $J_\rho(k) \in \mathbb{R}^{k\times k}$ has unit diagonal and coefficients equal to ρ everywhere else. $J_\rho(k)$ is a symmetric matrix, hence diagonalizable, and has eigenvalues $1-\rho$ with multiplicity $k-1$ and $1+(k-1)\rho$ with multiplicity one. As a consequence, we have, for $k \ge 2$,

$$\begin{aligned} \log\det \Sigma_S &= (k-1)\log(1-\rho) + \log(1+\rho(k-1)) \\ \mathrm{Tr}\,\Sigma_S^{-1} &= (n-k) + \frac{k-1}{1-\rho} + \frac{1}{1+\rho(k-1)}. \end{aligned}$$

The KL divergence is thus

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_S) \\
&\quad = \frac{1}{2}\Big[\mathrm{Tr}(\Sigma_1^{-1}\Sigma_0) - n - \log(\det\Sigma_0 / \det\Sigma_1)\Big] \\
&\quad = \frac{1}{2}\Big[(n-k) + \frac{k-1}{1-\rho} + \frac{1}{1+\rho(k-1)} - n + (k-1)\log(1-\rho) + \log(1+\rho(k-1))\Big] \\
&\quad = \frac{1}{2}\Big[k\Big(-1 + \frac{1}{1-\rho} + \log(1-\rho)\Big) - \Big(\frac{1}{1-\rho} + \log(1-\rho)\Big) \\
&\qquad\qquad + \Big(\frac{1}{1+\rho(k-1)} + \log(1+\rho(k-1))\Big)\Big].
\end{aligned}$$

□

5.6.2 Proof of Bound on KL Divergence

Proof. First note since the KL divergences are independent of n, it is sufficient to use the expressions of Proposition 23 with a contaminated set of size $s = |A \cap S| \le k$. As previously, we assume $s \ge 2$, as the result is trivial otherwise. Consider the expression for the KL divergence given in (5.9). Using (5.3), we obtain

$$\begin{aligned}
&\mathrm{KL}(\mathbb{P}_0|_A \| \mathbb{P}_S|_A) \\
&\quad = \mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_{S\cap A}) \\
&\quad \le \frac{1}{2}\Big[s\Big(-1 + \frac{1}{1-\rho} + \log(1-\rho) + \rho\Big) - \Big(\frac{1}{1-\rho} + \log(1-\rho)\Big) + \Big(\frac{1}{1+\rho} - \rho\Big)\Big] \\
&\quad = \frac{1}{2}\Big[s\Big(\rho + \frac{\rho}{1-\rho} + \log(1-\rho)\Big) + \frac{-2\rho}{1-\rho^2} - \log(1-\rho) - \rho\Big] \\
&\quad \le \frac{\rho s}{2(1-\rho)}.
\end{aligned}$$

Using (5.4) and (5.6), we obtain

$$\mathrm{KL}(\mathbb{P}_0 \| \mathbb{P}_S) \le \frac{1}{2}\Big[(s-1)^2\rho^2 + 2s\rho^2 + \rho^2\Big] = \frac{\rho^2}{2}\Big[(s-1)^2 + 2s + 1\Big] \le \frac{\rho^2 s(k+1)}{2}.$$

□

5.6.3 Proof of Proposition 18

Proof. We have $\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) = kf(\rho) + h(\rho)$ with

$$\begin{aligned}
f(\rho) &= \frac{1}{2}\Big[(1-\rho)^{-1} + \log(1-\rho) - 1\Big], \\
h(\rho) &= \frac{1}{2}\Big[-\Big(\frac{1}{1-\rho} + \log(1-\rho)\Big) + \Big(\frac{1}{1+(p-1)\rho} + \log(1+(p-1)\rho)\Big)\Big].
\end{aligned}$$

As previously, using (5.7), $f(\rho) \geq \frac{\rho^2}{16}$. Assume that $\rho k < 1$ and $k > 7$, then using (5.5) and (5.7),

$$\begin{aligned}
\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) &\geq \frac{\rho^2 k}{16} + h(\rho) \\
&\geq \frac{\rho^2 k}{16} - \frac{1}{2}\left[1 + 2\rho^2\right] + \frac{1}{2}\left[1 + \frac{\rho^2 (k-1)^2}{8}\right] \\
&= \rho^2 \left[\frac{k(k-1)^2}{16} - 1\right] \\
&\geq \frac{(\rho k)^2}{32}.
\end{aligned}$$

Now assume that $\rho k > 1$, then for $k > 32$,

$$\begin{aligned}
\mathrm{KL}(\mathbb{Q}_0 \| \mathbb{Q}_1) &\geq \frac{\rho^2 k}{16} - \frac{1}{2}\left[1 + 2\rho^2\right] + \frac{1}{2}\left[\frac{1}{1 + (k-1)\rho} + \log(1 + (k-1)\rho)\right] \\
&\geq \rho^2 \left[\frac{k}{16} - 1\right] + \frac{1}{2}\left[\frac{1}{1 + (k-1)\rho} + \log(1 + (k-1)\rho) - 1\right] \\
&\geq \frac{\rho^2 k}{32} + \frac{\log(1 + (k-1)\rho) - 1}{2}.
\end{aligned}$$

□

5.6.4 Proof of Proposition 19

Proof. With $p = \left\lceil \frac{1}{\rho} \right\rceil$, when $\rho k > 1$, we have $\left\lceil \frac{1}{\rho} \right\rceil < k + 1$, and as a consequence,

$$\mathrm{KL}(\mathbb{P}_0^p \| \mathbb{P}_S^p) \geq \frac{\log 2 - 1/2}{2} \geq \frac{1}{11}.$$

□

5.7 Appendix: Extensions to Unnormalized Model

5.7.1 Uniform (non-adaptive) Lower Bound

Proposition 24. *For any class $\mathcal{C}$, any $\rho \in [0, 0.9)$, the minimum risk in the normalized model (resp. the unnormalized model) under uniform (non-adaptive) sensing is bounded as*

$$R^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{E\left[\cosh^m\left(\frac{8\rho Z}{1-\rho}\right)\right] - 1}$$
$$\bar{R}^* \geq \frac{1}{2} - \frac{1}{4}\sqrt{E\left[\cosh^m\left(8\rho Z\right)\right] - 1}$$

where Z is the size of the intersection of two elements of $\mathcal{C}$ drawn independently and uniformly at random.

Proof. This is essentially a reproduction of the proof of (Arias-Castro et al., 2014) with minor modifications. The details are omitted. □

5.7.2 Uniform (non-adaptive) Upper Bound

Let $H(b) = b - 1 - \log b$ for $b > 1$.

Proposition 25. *Under uniform (non-adaptive) sensing, the localized square-sum test that rejects when*

$$Y_{scan} = \max_{S \in \mathcal{C}} \sum_{t=1}^{m} \left(\sum_{i \in S} X_i^t \right)^2$$

exceeds

$$\frac{1}{2}\left(\rho k^2 m + H^{-1}(3 \log |\mathcal{C}|/m) - 1)km\right)$$

is asymptotically powerful when

$$\rho k \geq c_1 \max\left(\sqrt{\frac{\log |\mathcal{C}|}{m}}, \frac{\log |\mathcal{C}|}{m}\right)$$

both for the normalized and unnormalized models.

Proof. This is proved in (Arias-Castro et al., 2014) for the normalized model. In the case of the unnormalized model, the test statistic is distributed as $k\chi^2_m$ under the null, and as $(k(1+\rho)+\rho k(k-1))\chi^2_m$ under the alternative, which changes only mildly the proof with respect to the normalized model. □

5.7.3 KL Divergences

Proposition 26. *We have*

$$\mathrm{KL}(\bar{\mathbb{P}}_0 \| \bar{\mathbb{P}}_S) = \frac{\mathbf{1}_{k \geq 2}}{2}\left[-1 + \frac{1}{1+\rho k} + \log(1+\rho k)\right]. \tag{5.10}$$

Proof. The KL divergence between $\bar{\mathbb{P}}_0$ and $\bar{\mathbb{P}}_S$ can be computed using the standard formula for KL divergence between two centered Gaussian vectors, with covariances matrices

$$\Sigma_0 = I_n, \quad \Sigma_1 = \bar{\Sigma}_S.$$

When $k = 0$, the divergence is zero, and we will thus assume $k \geq 1$. Up to a simultaneous permutation of rows and columns,

$$\bar{\Sigma}_S = \begin{bmatrix} I_{n-k} & \\ & I_k + K_\rho(k) \end{bmatrix}$$

where $K_\rho(k) \in \mathbb{R}^{k\times k}$ has coefficients equal to ρ everywhere. Like previously, $I_k + K_\rho(k)$ is diagonalizable, and has eigenvalue 1 with multiplicity $k-1$, and eigenvalue $1+\rho k$ with multiplicity one. As a consequence, for $k \geq 1$, we have

$$\log\det\bar{\Sigma}_S = \log(1+\rho k)$$
$$\operatorname{Tr}\bar{\Sigma}_S^{-1} = (n-1) + \frac{1}{1+\rho k}.$$

This leads to

$$\begin{aligned}\mathrm{KL}(\bar{\mathbb{P}}_0\|\bar{\mathbb{P}}_S) &= \frac{1}{2}\Big[\operatorname{Tr}(\Sigma_1^{-1}\Sigma_0) - n - \log(\det\Sigma_0/\det\Sigma_1)\Big] \\ &= \frac{1}{2}\left[(n-1) - n + \frac{1}{1+\rho k} + \log(1+\rho k)\right].\end{aligned}$$

□

Proposition 27. *For any $A \subset [n]$,*

$$\mathrm{KL}(\bar{\mathbb{P}}_0|_A\|\bar{\mathbb{P}}_S|_A) \leq \min\left[\frac{\rho}{2}, \frac{\rho^2 k}{2}\right]|A\cap S|.$$

Proof. First note since the KL divergences are independent of n, it is sufficient to use the expressions of Proposition 23 with a contaminated set of size $s = |A\cap S|$. As previously, we assume $s \geq 1$, as the result is trivial otherwise. Consider the unnormalized model, with KL divergence given in (5.10). Using (5.3), we obtain

$$\mathrm{KL}(\bar{\mathbb{P}}_0|_A\|\bar{\mathbb{P}}_S|_A) = \mathrm{KL}(\bar{\mathbb{P}}_0\|\bar{\mathbb{P}}_{A\cap S}) \leq \frac{\rho s}{2}.$$

Using (5.4) we obtain

$$\mathrm{KL}(\bar{\mathbb{P}}_0|_A\|\bar{\mathbb{P}}_S|_A) = \mathrm{KL}(\bar{\mathbb{P}}_0\|\bar{\mathbb{P}}_{A\cap S}) \leq \frac{\rho^2 s^2}{2} \leq \frac{\rho^2 sk}{2}.$$

Combining these last two inequalities yields the desired result. □

Proposition 28. *Assume that ρ converges to zero, and that either*

$$\rho k \to \infty \quad \textit{and} \quad m\log(1+\rho k) \geq C_3\log\log(n/k),$$

or

$$\rho k \to 0 \quad \textit{and} \quad \rho k\sqrt{m} \geq C_4\sqrt{\log\log(n/k)}$$

for some constants C_3 and C_4. Then sequential thresholding for disjoint k-intervals has risk converging to zero.

Proof. For the unnormalized model, when $\rho k > 1$, using (5.8),

$$\mathrm{KL}(\bar{\mathbb{Q}}_0 \| \bar{\mathbb{Q}}_1) \geq \frac{\log(1+\rho k)}{10}.$$

When $\rho k < 1$, using (5.7),

$$\mathrm{KL}(\bar{\mathbb{Q}}_0 \| \bar{\mathbb{Q}}_1) \geq \frac{(\rho k)^2}{16}.$$

□

Proposition 29. *Assume that ρ converges to zero, and that either*

$$\rho k \to \infty \quad \text{and} \quad \rho k m \geq C_5 \log\log(n/k),$$

or

$$\rho k \to 0 \quad \text{and} \quad \rho k \sqrt{m} \geq C_6 \sqrt{\log\log(n/k)}$$

for some constants C_5, C_6. Then the modified sequential thresholding for disjoint k-intervals has risk converging to zero.

Proof. For the unnormalized model with $p = \left\lceil \frac{1}{\rho} \right\rceil$, when $\rho k > 1$, we have $\left\lceil \frac{1}{\rho} \right\rceil < k+1$, and as a consequence,

$$\mathrm{KL}(\bar{\mathbb{P}}_0^p \| \bar{\mathbb{P}}_S^p) \geq \frac{\log 2 - 1/2}{2} \geq \frac{1}{11}.$$

□

6

Detection of Objects with High-dimensional CNN Features

"I well recall a group of us, after a session on the IBM 701 at a meeting where they talked about the proposed 18 machines, all believed this would saturate the market for many years!"

— Richard Hamming

Contents

6.1 Introduction . . . 132
6.2 Detection with DPMs . . . 133
6.2.1 Detection Task . . . 133
6.2.2 Deformable Part Models . . . 133
6.3 Integrating Convolutional Features into DPMs . . . 135
6.3.1 The Alexnet Network Structure . . . 135
6.3.2 Prior Work . . . 136
6.3.3 Using CNN Layer 5 Features in DPMs . . . 137
6.4 Results on Pascal VOC 2007 . . . 138

6.1 Introduction

In the object detection problem, the goal is to find whether certain types of objects are present in an image, and to precisely estimate their positions when this is the case. In the last decade, multiple datasets of images with object annotations and bounding boxes have appeared, including the subsequent PASCAL VOC datasets (Everingham et al., 2007, 2010a, 2012), the INRIA Person dataset (Dalal and Triggs, 2005), or the the ILSVRC datasets based on subsets of the ImageNet dataset. This has bolstered the development of detection methods significantly, not simply due to the availability of clean data, but also as it provides a common and reproducible framework for comparing algorithms and methods.

In recent years, the detection tasks have been dominated by structured methods (Felzenszwalb et al., 2010b) such as deformable part models (DPMs). DPMs are a particular instance of *sliding window* detector. In DPMs, an object is modeled as a tree of parts. Both the appearance and the usual relative positions of the parts are modeled, which allows to be robust to shape deformations. In order to further account for object variabilities, mixture models can be trained for each object type, e.g., using latent SVMs classifiers (Felzenszwalb et al., 2010b). Through this structure, DPMs can be extended to address much more general problems such as human pose estimation (Yang and Ramanan, 2011), facial expression recognition (Zhu and Ramanan, 2012b), or three-dimensional structure estimation (Kakadiaris et al., 2007). As many computer vision models, DPMs are based on hand-crafted image features. In computer vision, features are usually designed to encode color or geometrical properties of images while enforcing various type of invariances (e.g., scale invariance with SIFT or HOG). This prompts numerous model selection questions and has generated a very large body of work.

Convolutional neural networks (CNNs) (Hinton and Osindero, 2006; Ranzato et al., 2011; Krizhevsky et al., 2012; Sermanet et al., 2013; Farabet et al., 2013) have recently appeared to be particularly effective at the image recognition task, as was shown on the ImageNet dataset during the latest ILSVRC annual image classification and object detection competitions (Deng et al., 2009). CNNs are variants of multilayer perceptrons with a sparse and local connectivity between neurons of successive layers. CNNs and coding techniques (such as sparse coding, or structured coding methods) consider a largely different stand from hand-crafted features: the representation is to be learned from the data, either in an unsupervised fashion (auto encoders, and coding), or coupled with a specific pattern recognition task. The main challenge with such techniques is to make the best use of their flexibility, while staying in control of their computational complexity. Many computer vision models have already been revisited using these *representation learning* techniques (Eslami et al., 2012; Tang et al., 2012), and one can anticipate this will only intensify in the future.

In spite of this, it is still not clear how to exploit representation learning techniques for fast and accurate localization of multiple objects in images. Most existing validations of such models have focused on classification or labeling tasks, wherein the objective consists only in identifying a dominant object type in an image. On the other hand, the detection problem is a *structured prediction* problems where a label is not only an object type, but also the coordinates of the bounding box where the object is located. Recently, there have been attempts at addressing this problem of using CNNs for detection (Girshick et al., 2014; Sermanet et al., 2014; Iandola et al., 2014), albeit not in a structured framework such as DPMs.

In this work, we demonstrate that using CNN feature pyramids within DPMs allows to obtain a significant boost in DPM performance on VOC2007. In Section 6.2, we review the object detection problem and the DPM framework. In Section 6.3, we look at CNNs, and describe how such feature representations can be integrated with DPMs. Finally, we present experimental results on VOC2007 in Section 6.4, and we close this chapter with a discussion of the results in Section **??**. This chapter is joint work with Iasonas Kokkinos and Stavros Tsogkas.

6.2 Detection with DPMs

In this section, we review the detection task, and give a high-level overview of DPMs.

6.2.1 Detection Task

In computer vision, the detection task consists in both identifying the types of the objects in an image, and precisely localizing them. Detection results are usually evaluated based on the following measure of overlap between a predicted bounding box, and an annotated bounding box. The Intersection over Union (`IoU`, also referred to as *Jaccard's coefficient*) of two bounding boxes $B_1 = \{x_1^{\min}, \ldots, x_1^{\max}\} \times \{y_1^{\min}, \ldots, y_1^{\max}\}$ and $B_2 = \{x_2^{\min}, \ldots, x_2^{\max}\} \times \{y_2^{\min}, \ldots, y_2^{\max}\}$ is

$$\texttt{IoU}(B_1, B_2) = \frac{|B_1 \cap B_2|}{|B_1 \cup B_2|}.$$

In the PASCAL VOC detection challenges (Everingham et al., 2007), a predicted bounding box for a given class is considered a true positive if the best `IoU` with a ground truth bounding box for this class exceeds 0.5. A given ground truth bounding box can only be used for a single predicted bounding box, such that predicting many close bounding boxes leads to many false positives. Predictions are required to be accompanied by a confidence score, such that a varying threshold can be used to explore different detection regimes. The final evaluation is in terms of precision/recall curve, which is usually summarized with the average precision (AP, the area under the precision/recall curve). The high recall regime corresponds to many bounding box predictions thus capturing all the objects, while the high precision regime corresponds to a small number of relevant bound box predictions.

Understanding failures of detectors is paramount to improve their performance. In detection, errors can be of different types. Localization errors correspond to where an object was correctly detected, but poorly localized so that the `IoU` does not exceed the threshold (this includes duplicate detections). False positives can be classified (Hoiem et al., 2012) into confusion with similar objects, confusion with dissimilar objects, or confusion with background. This is illustrated in Figure 6.1.

6.2.2 Deformable Part Models

In this section, we review star-shaped deformable part models. These models are defined by a root node which models the global appearance of the object, and part nodes which are allowed to float around the position of the root node to model certain specific features of the object. As a simple example, the problem of human pose estimation can be cast as such a star-shaped DPM, with parts for the head, the torso, the arms, and the legs. A given bounding box is scored in DPMs according to the score of the root model, of the parts

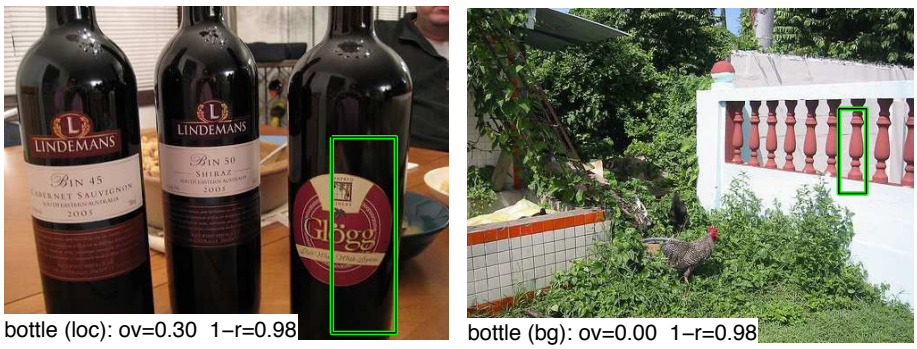

Figure 6.1: Localization error (left), confusion with background (right)

models, and of the value of interaction terms between the root and the parts. When training in the absence of precise annotation for the parts, their positions are treated as latent variables, and are estimated. When doing detection, the score of a bounding box corresponds to the highest scoring configuration over all possible parts locations. Examples of detections along with the latent part locations are shown in Figure 6.2.

Figure 6.2: Part based models, sample detections from Felzenszwalb et al. (2010b): detection bounding box (red), latent parts (blue)

Formally, DPMs associate linear filters with the root and the part nodes. The part filters are usually square filters of fixed size, while the root filters are larger. Consider a root-part configuration $z = (z_0, z_1, \ldots z_P)$, where z_0 is the image location of the center of the root filter, while, $z_1, \ldots, z_P$ are the image locations of the centers of each of the part filters. Denote by $I(z)$ a feature representation of the image patch around image location z, of size corresponding to that of the root filter or of the part filters, depending on context. A configuration is scored using a model of the form

$$M(z) = \langle w_0, I(z_0) \rangle + \sum_{p=1}^{P} \Big[\langle w_p, I(z_p) \rangle + B_p(z_p, z_0) \Big].$$

Here, $w_0, w_1, \ldots, w_P$ are the filters (in feature space), and the dot product describes how well the patch at this position matches the part model. The terms B_p are the *second order potentials*, and depend only on the relative position difference $z_p - z_0$. For instance, in the case of human pose estimation, one may want to penalize when the head part is not

strictly above the torso, or when the left arm is not to the left of the torso. These terms are usually simple quadratic forms. At training time, both the filters $w_1, \ldots, w_P$ and the second order potentials $B_1, \ldots, B_P$ must be learned.

Using this model, bounding box predictions can be produced as follows. First, we decide on an aspect ratio and a putative root filter location $\widehat{z_0}$, that leads to a candidate bounding box. The best score over all compatible part configurations is

$$S(\widehat{z_0}) = \max_{z : z_0 = \widehat{z_0}} M(z).$$

This provides a confidence score. Finally, this procedure is actually used for all possible root positions $\widehat{z_0}$, which justifies that DPMs are referred to as *dense* detectors. For this reason, DPMs may be slow both to train and to evaluate. However, DPMs can be accelerated significantly at test time using cascades (Felzenszwalb et al., 2010a), branch and bound (Kokkinos, 2011), or score approximations (Kokkinos, 2013).

The procedure that we have described can be used to detect objects at a fixed scale that depends on the size of the part and of the root filters. In order to detect objects at all scales, a *feature pyramid* is used instead of a simple single-scale feature representation of the image. At an elementary level, this is equivalent to performing detection on rescaled version of a single image. This is illustrated on Figure 6.3: down-sampled versions of the image may allow to detect the head of the horse, while higher resolutions may allow to detect finer details such as the ears of the horse. In addition, a mixture of such models can be considered, leading to an additional latent variable which is the mixture component identifier. These mixture components allow to account for different viewpoints or subtypes of objects within a given class. In particular, root filters with different aspect ratios are traditionally used.

6.3 Integrating Convolutional Features into DPMs

In this chapter, we consider the core problem of this chapter: how can we use features from CNNs for detection with DPMs, instead of using HOG features or similar hand-crafted representations? In the remainder of this section, we go into detail into the structure of the CNN that we consider, as well as into practical details for integrating deep features into DPMs.

6.3.1 The Alexnet Network Structure

Most of recent works with CNNs for image classification or object detection are focused on the *Alexnet* network (Krizhevsky et al., 2012), which has won recent ImageNet object classification challenges. The *Alexnet* name is used both to refer to the network structure (i.e., the number, shapes and types of layers), and to refer to already-trained networks based on this structure. The first five layers are convolutional, and consist only of convolutions and pooling layers. Hence, the first five layers are intrinsically translation equivariant. The last two layers are classical fully connected layers as in multi-layer perceptrons. The input of the network consists of a $224 \times 224 \times 3$ patch, which is transformed to two $13 \times 13 \times 128$ patches after layer 5, and to a 4096 dimensional vector after layer 7. While the features obtained at the convolutional layers have a spatial interpretation (as is the case with HOG), this is not the case at the fully-connected layers.

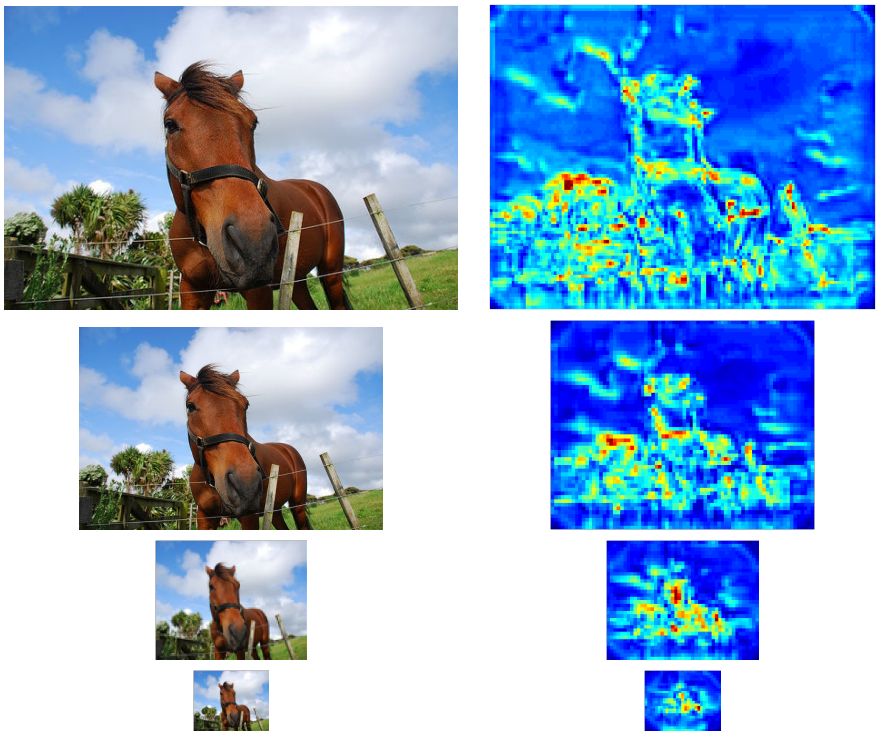

Figure 6.3: Feature pyramid: original image at different scales (left), energy of CNN feature representation at layer 5 (right)

The Alexnet network was trained on the ImageNet database, where all images were rescaled to a fixed size, and the mean of all these images was subtracted. In order to augment the data for training, random translations and horizontal reflections were used, and have been observed to reduce over-fitting. However, the Alexnet network was trained for classification, and it may be beneficial to retrain the network when working on a related but different task such as detection, as was observed for R-CNN, where this is referred to as *fine-tuning* (Girshick et al., 2014). This can be achieved by setting up a supervised objective adapted to the task at hand, and performing back-propagation on the original network.

6.3.2 Prior Work

An interesting attempt at using CNNs for detection has been made with OverFeat (Sermanet et al., 2014), which considers a sliding window approach. Two CNNs are jointly trained to predict the class of the object and the coordinates of the bounding box containing the object, respectively, from a given input window. These networks are then fed with all possible windows from the original images, at different scales. A squared loss is used as the loss for the bounding box regression problem.

The *Regions with CNN features* (R-CNN) method (Girshick et al., 2014) addresses the detection problem in a remarkably simple and effective way, through the use of region proposals (Uijlings et al., 2013). Many promising regions are obtained from segmentation considerations and warped to a fixed size window, which is then used as input to a CNN. This is essentially a reduction to a classification task, but performs extremely well, pro-

viding a mean average precision (mAP) of 58.5% on 20-classes VOC2007, and of 31.4% on 200-classes ILSVRC2013. These large mAPs are at the expense, however, of the time spent training and testing: even though the method scales well with the number of classes, the computation of the CNN features over all warped region proposals is expensive, even on modern GPU machines. In addition, R-CNN is not a structured detector, and hence cannot be generalized easily to more difficult structure estimation problems.

6.3.3 Using CNN Layer 5 Features in DPMs

Due to translational equivariance, the output of the convolutional layers can actually be computed on images larger than the input patch size used for training. This was leveraged in OverFeat (Sermanet et al., 2014) and in other recent works (Iandola et al., 2014; He et al., 2014). In particular, Iandola et al. (2014) proposed to compute a feature pyramid based on the convolutional layers (i.e., using the layer 5 features of the network). In order to obtain a multi-scale representation, the patchwork of scales approach (Dubout and Fleuret, 2012) is used: different scales of an image are stitched in a large patchwork image as shown in Figure 6.4, and the feature representation of the patchwork image is computed and transformed back into a pyramid. This work demonstrates how to compute feature pyramids based on a convolutional network, but no quantitative evaluation on a specific task is provided.

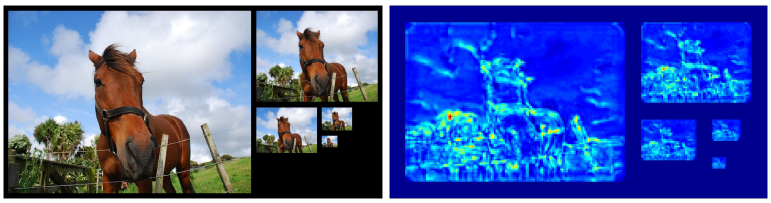

Figure 6.4: Patchwork method: original stitched image (left), energy of CNN feature representation at layer 5 (right)

However, a full image cannot be directly propagated through the fully-connected layers, which require a fixed size input that has to be fixed at the time where the network is trained. From a feature pyramid at layer 5, one may crop (or warp to) patches of fixed size to compute the output of the fully-connected layers (He et al., 2014), or simply stick to the spatially arranged features at layer 5. This is in contrast with the approach used in R-CNN which first warps to a fixed size before even the convolutional layer. In order to have a structured detector like with DPMs, we need to conserve a spatial arrangement of the features, as this allows to share computations when scoring part configurations. As a consequence, we choose in this work to directly use the feature pyramid based directly upon the layer 5 features.

When extracting such features, a key quantity is the subsampling factor `sub` from the original input to the spatial feature representation, also referred to as the bin size, as illustrated in Figure 6.5. For Alexnet, `sub` $= 16$, such that each feature is computed over 16×16 cells over the input image. As this corresponds to large bins, we oversampled by a factor of two all images before computing features, which effectively leads to `sub` $= 8$. We note that in Figure 6.3, the features have been resized by a factor of `sub` in order to appear of the same size as the images. Figure 6.5 illustrates this with `sub` $= 16$: each pixel in the

feature representation corresponds to a 16×16 cell in the original image. Higher resolu-

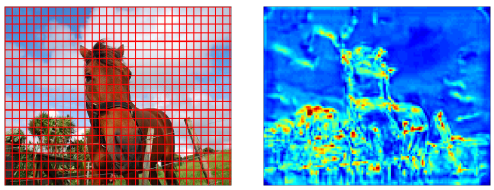

Figure 6.5: Illustration of subsampling factor

tions can be obtained similarly, e.g., $\texttt{sub} = 4$ can be obtained by oversampling by a factor of four, although feature computation as well as training and testing become prohibitively long. The layer 5 features in each bin are 256-dimensional, which is eight times more than with HOG features, which further increases the computational challenges associated with training such models.

With HOG, the features can be interpreted as normalized frequencies for orientations of the gradients, and can thus be represented visually, or even transformed back into the image domain to some extent (Vondrick et al., 2013). With layer 5 CNN features, this type of interpretation is not possible. Although we have represented CNN features through their total energy so far, it would be desirable to have a way to visualize the feature representations of the images, as well as the filters.

6.4 Results on Pascal VOC 2007

Instead of the Alexnet network, we use the R-CNN network fine-tuned for detection with region proposals. We only report results with $\texttt{sub} = 8$, as $\texttt{sub} = 16$ leads to significantly worse mAPs, while $\texttt{sub} = 4$ is computationally prohibitive. We consider a normalization for the CNN feature that is based on the **L2-Hys** (for *hysteresis*) scheme (Dalal and Triggs, 2005). Given a feature vector x, normalize in Euclidean norm:

$$x \leftarrow \frac{x}{\|x\|_2 + \varepsilon},$$

with $\varepsilon = 10^{-3}$. Then, x is capped componentwise to a maximum value $v = 0.1$:

$$x \leftarrow \min(x, v).$$

We note that the layer 5 features are nonnegative, such that this is equivalent to capping to a maximum amplitude. Finally, x is normalized again in Euclidean norm. In addition to **L2H-Hys**, we investigated other normalization schemes, such as simple Euclidean norm normalization, or ℓ_1-norm normalization. Both of these alternative schemes yielded similar results.

In DPMs, it is common to train distinct mixture components for left-facing and right-facing objects. When using HOG features, one may avoid to double the number of linear classifiers by leveraging the fact that the HOG features of a flipped patch can be obtained by shuffling the HOG features of the original patch. This is because HOG features are defined in terms of the geometrical orientations of the image gradient. Hence, given a filter for a left-facing object, one may obtain in closed-form a filter for right-facing objects

through this shuffling process. This is no longer the case for CNN features. Instead, for each image, we compute two feature pyramids, one on the original image, and one on the mirror image. For each left-facing filter, we define a right-facing filter which is constrained to be the mirror of the left-facing filter. However, unlike the left-facing filter, the right-facing filter operates on the feature pyramid of the mirror image.

Due to the non-convexity of the latent SVMs models, initialization has to be performed with care. Recently, a simpler method (Girshick and Malik, 2013) was proposed based on Linear Discriminant Analysis. This method was shown to yield similar mAPs to the full-blown DPM pipeline, and was used throughout all our experiments.

Our results are reported in Table 6.1. We consider two variants of our method: the first one, C-DPM, combines sliding window detection followed by nonmaximum suppression; the second one, C-DPM-BB, is augmented with bounding box regression, using the original bounding box coordinates as input features. We compare these two variants to the following methods: DPMv5 refers to the baseline DPM implementation using HOG features and bounding-box regression, as in (Felzenszwalb et al., 2010b), while RCNN5, RCNN7, RCNN7-BB correspond to the performance of (fine-tuned) RCNN using layer 5 features, layer 7 features, or layer 7 features with an extra bounding box regression based on (richer) CNN features, respectively. The last rows of the second and third blocks indicate the difference between the AP achieved by our method and DPMv5 or RCNN5, respectively. In order to obtain commensurate performance measures, we compare DPMv5 with our variant that includes bounding box regression (C-DPM-BB), and RCNN5, which does not include bounding box regression, to C-DPM.

From the second block of Table 6.1, it is clear that we significantly improve over HOG-based DPMs, while employing the exact same training pipeline; this is indicating the clear boost we obtain simply by changing the low-level image features. However, the results are not as clear-cut when it comes to comparing with RCNN. Even when comparing only to RCNN-5, we have a moderate drop in performance, while our DPMs are still quite behind RCNN-7. The difference with respect to RCNN-7 can be attributed to the better discriminative power of deeper features and could be addressed by incorporating nonlinear classifiers, or computing all features up to layer 7 in a convolutional manner.

Method	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	dtbl	dog	hors	mbike	person	plant	sheep	sofa	train	tv	mAP
C-DPM	39.7	59.5	35.8	24.8	35.5	53.7	48.6	46.0	29.2	36.8	45.5	42.0	57.7	56.0	37.4	30.1	31.1	50.4	56.1	51.6	43.4
C-DPM-BB	50.9	64.4	43.4	29.8	40.3	56.9	58.6	46.3	33.3	40.5	47.3	43.4	65.2	60.5	42.2	31.4	35.2	54.5	61.6	58.6	48.2
DPMv5	33.2	60.3	10.2	16.1	27.3	54.3	58.2	23.0	20.0	24.1	26.7	12.7	58.1	48.2	43.2	12.0	21.1	36.1	46.0	43.5	33.7
C-DPM-BB vs. DPMv5	+17.7	+4.1	+33.2	+13.7	+13.0	+2.6	+0.4	+23.3	+13.3	+16.4	+20.6	+30.7	+7.1	+12.3	-1.0	+19.4	+14.1	+18.4	+15.6	+15.1	+14.5
RCNN7-BB	68.1	72.8	56.8	43.0	36.8	66.3	74.2	67.6	34.4	63.5	54.5	61.2	69.1	68.6	58.7	33.4	62.9	51.1	62.5	64.8	58.5
RCNN7	64.2	69.7	50.0	41.9	32.0	62.6	71.0	60.7	32.7	58.5	46.5	56.1	60.6	66.8	54.2	31.5	52.8	48.9	57.9	64.7	54.2
RCNN5	58.2	63.3	37.9	27.6	26.1	54.1	66.9	51.4	26.7	55.5	43.4	43.1	57.7	59.0	45.8	28.1	50.8	40.6	53.1	56.4	47.3
C-DPM vs. RCNN5	-18.5	-3.8	-2.1	-2.8	+9.4	-0.4	-18.3	-5.4	+2.5	-18.7	+2.1	-1.1	0.0	-3.0	-8.4	+2.0	-19.7	+9.8	+3.0	-4.8	-3.9

Table 6.1: Results on PASCAL VOC 2007: average precision in percent

A intriguing point is the difference in performance between RCNN-5 and C-DPM, since both use the same features. One would expect DPMs to have better performance (since they do not rely on region proposals, and also come with many mixtures and deformable parts), but this is not the case. We suspect that this is because (i) DPMs split the training set into roughly 3 subsets (for the different aspect ratios/mixtures), effectively reducing by 3 the amount of training data and (ii) DPMs are somewhat rigid when it comes to the kind of aspect ratio that they can deal with, (3 fixed ratios) which may be problematic in

the presence of large aspect ratio variatios; by contrast RCNN warps all region proposals images onto a single canonical scale. To conclude, we have shown that replacing HOG features with CNN features yields a substantial improvement in DPM detection performance; given the widespread use of DPMs in a broad range of structured prediction tasks (Yang and Ramanan, 2013; Zhu and Ramanan, 2012a), we anticipate that this will soon become common practice.

Bibliography

L. Addario-Berry, N. Broutin, L. Devroye, and G. Lugosi. On combinatorial testing problems. *Annals of Statistics*, 38:3063–3092, 2010.

A. Agarwal, S. Negahban, M. J. Wainwright, et al. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40(2):1171–1197, 2012.

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

F. Alizadeh and D. Goldfarb. Second-order cone programming. *Mathematical Programming*, 95(1):3–51, 2003.

N. Alon and A. Naor. Approximating the cut-norm via grothendieck's inequality. *SIAM Journal on Computing*, 35(4):787–803, 2006.

D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp. Living on the edge: A geometric theory of phase transitions in convex optimization. *arXiv preprint arXiv:1303.6672*, 2013.

B. P. Ames and S. A. Vavasis. Nuclear norm minimization for the planted clique and biclique problems. *Mathematical programming*, 129(1):69–89, 2011.

A. A. Amini and M. J. Wainwright. High-dimensional analysis of semidefinite relaxations for sparse principal components. In *Information Theory, 2008. ISIT 2008. IEEE International Symposium on*, pages 2454–2458. IEEE, 2008.

A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *arXiv preprint arXiv:1210.7559*, 2012.

A. Argyriou, C. A. Micchelli, and M. Pontil. When is there a representer theorem? vector versus matrix regularizers. *Journal of Machine Learning Research*, 10:2507–2529, 2009.

A. Argyriou, R. Foygel, and N. Srebro. Sparse prediction with the k-support norm. In *Proc. NIPS*, pages 1457–1465, 2012.

E. Arias-Castro, E. J. Candès, H. Helgason, and O. Zeitouni. Searching for a trail of evidence in a maze. *Annals of Statistics*, 36:1726–1757, 2008.

E. Arias-Castro, S. Bubeck, and G. Lugosi. Detection of correlations. *Annals of Statistics*, 40(1):412–435, 2012.

E. Arias-Castro, E. J. Candes, and M. A. Davenport. On the fundamental limits of adaptive sensing. *IEEE Trans. Information Theory*, 59(1):472–481, 2013.

E. Arias-Castro, S. Bubeck, and G. Lugosi. Detecting positive correlations in a multivariate sample. *Bernoulli*, 2014.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple Kernel Learning, Conic Duality, and the SMO Algorithm. In *Proc. ICML*, 2004.

F. R. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. *Proc. NIPS*, 17:73–80, 2005.

G. Bakir, T. Hofmann, B. Scholkopf, A. Smola, B. Taskar, and S. Vishanathan. *Predicting structured data*. MIT Press, 2007.

O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine Learning Research*, 9:485–516, 2008.

Y. Baraud. Non-asymptotic minimax rates of testing in signal detection. *Bernoulli*, 8:577–606, 2002.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2003.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. PAMI*, 24(4):509–522, 2002.

S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In *Proc. NIPS*, pages 163–171, 2010.

R. Berinde, A. C. Gilbert, P. Indyk, H. Karloff, and M. J. Strauss. Combining geometry and combinatorics: A unified approach to sparse signal recovery. In *Communication, Control, and Computing, 2008 46th Annual Allerton Conference on*, pages 798–805. IEEE, 2008.

Q. Berthet and P. Rigollet. Optimal detection of sparse principal components in high dimension. *Annals of Statistics*, 41(1):1780–1815, 2013.

D. Bertsekas. Incremental gradient, subgradient, and proximal methods for convex optimization: a survey. *Optimization for Machine Learning*, page 85, 2011.

J. Bi, T. Zhang, and K. P. Bennett. Column-generation boosting methods for mixture of kernels. In *Proc. KDD*, pages 521–526. ACM, 2004.

J. Bien and R. Tibshirani. Sparse estimation of a covariance matrix. *Biometrika*, 2010.

A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *Proc. COLT*, pages 92–100. ACM, 1998.

A. Blum, G. Konjevod, R. Ravi, and S. Vempala. Semi-definite relaxations for minimum bandwidth and other vertex-ordering problems. In *Proceedings of the thirtieth annual ACM Symposium on Theory of computing*, pages 100–105. ACM, 1998.

T. Blumensath and M. E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

J. Bock and D. Gough. Predicting protein–protein interactions from primary structure. *Bioinformatics*, 17(5):455–460, 2001.

B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. COLT*, pages 144–152. ACM, 1992.

J. Bourgain. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52(1-2):46–52, 1985.

P. Bühlmann and S. Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer, 2011.

C. Butucea, Y. I. Ingster, et al. Detection of a sparse submatrix of a high-dimensional noisy matrix. *Bernoulli*, 19(5B):2652–2688, 2013.

J. Cai, E. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM J. on Optimization*, 20(4):1956–1982, 2008.

T. T. Cai and W. Zhou. Matrix completion via max-norm constrained optimization. *arXiv preprint arXiv:1303.0341*, 2013.

T. T. Cai, C.-H. Zhang, H. H. Zhou, et al. Optimal rates of convergence for covariance matrix estimation. *Annals of Statistics*, 38(4):2118–2144, 2010.

T. Cai, Z. Ma, and Y. Wu. Optimal estimation and rank detection for sparse spiked covariance matrices. *Probability Theory and Related Fields*, pages 1–35, 2013.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

L. Cao, J. Luo, F. Liang, and T. S. Huang. Heterogeneous feature machines for visual recognition. In *Proc. ICCV*, pages 1095–1102. IEEE, 2009.

R. M. Castro. Adaptive sensing performance lower bounds for sparse signal estimation and testing. *arXiv preprint arXiv:1206.0648*, 2012.

R. M. Castro, G. Lugosi, and P.-A. Savalle. Detection of correlations with adaptive sensing. *arXiv preprint arXiv:1311.5366*, 2013.

V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Opt.*, 21:572–596, 2011.

V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

C.-C. Chang and C.-J. Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

S. Chatterjee, S. Chen, and A. Banerjee. Generalized dantzig selector: Application to the k-support norm. *arXiv preprint arXiv:1406.5291*, 2014.

G. Chechik, A. Globerson, N. Tishby, and Y. Weiss. Information bottleneck for gaussian variables. *Journal of Machine Learning Research*, pages 165–188, 2005.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

S. Chen and D. Donoho. Basis pursuit. In *Signals, Systems and Computers, 1994. 1994 Conference Record of the Twenty-Eighth Asilomar Conference on*, volume 1, pages 41–44. IEEE, 1994.

Y. Chen and A. Krause. Near-optimal batch mode active learning and adaptive submodular optimization. In *Proc. ICML*, 2013.

A. Cichocki, R. Zdunek, A. H. Phan, and S.-i. Amari. *Nonnegative matrix and tensor factorizations: applications to exploratory multi-way data analysis and blind source separation.* John Wiley & Sons, 2009.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the American Mathematical Society*, 22(1):211–231, 2009.

P. Combettes and J. Pesquet. Proximal splitting methods in signal processing. *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212, 2011.

T. M. Cover and J. A. Thomas. *Elements of information theory.* John Wiley & Sons, 2012.

E. Cuthill and J. McKee. Reducing the bandwidth of sparse symmetric matrices. In *Proceedings of the 1969 24th ACM National Conference*, pages 157–172. ACM, 1969.

N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proc. CVPR*, volume 1, pages 886–893. IEEE, 2005.

A. S. Dalalyan and A. B. Tsybakov. Aggregation by exponential weighting and sharp oracle inequalities. In *Proc. COLT*, pages 97–111. Springer, 2007.

S. Dasgupta and A. Gupta. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2003.

A. d'Aspremont, L. El Ghaoui, M. Jordan, and G. Lanckriet. A direct formulation for sparse pca using semidefinite programming. *SIAM review*, 49(3):434–448, 2007.

A. d'Aspremont and L. El Ghaoui. Testing the nullspace property using semidefinite programming. *Mathematical programming*, 127(1):123–144, 2011.

A. Demiriz, K. P. Bennett, and J. Shawe-Taylor. Linear programming boosting via column generation. *Machine Learning*, 46(1-3):225–254, 2002.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, pages 248–255. IEEE, 2009.

D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Annals of Statistics*, 32:962–994, 2004.

D. L. Donoho and J. Tanner. Sparse nonnegative solution of underdetermined linear equations by linear programming. *Proceedings of the National Academy of Sciences of the United States of America*, 102(27):9446–9451, 2005.

D. L. Donoho and J. Tanner. Counting the faces of randomly-projected hypercubes and orthants, with applications. *Discrete & computational geometry*, 43(3):522–541, 2010.

D. Drusvyatskiy, S. Vavasis, and H. Wolkowicz. Extreme point inequalities and geometry of the rank sparsity ball. *arXiv preprint arXiv:1401.4774*, 2014.

C. Dubout and F. Fleuret. Exact acceleration of linear object detectors. In *Proc. ECCV*, pages 301–311. Springer, 2012.

J. Dunagan and S. Vempala. On euclidean embeddings and bandwidth minimization. In *Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques*, pages 229–240. Springer, 2001.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

N. El Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, 2009.

S. M. A. Eslami, N. Heess, and J. M. Winn. The shape boltzmann machine: A strong model of object shape. In *Proc. CVPR*, 2012.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html, 2007.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html, 2010a.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html, 2012.

M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *International journal of Computer Vision*, 88(2):303–338, 2010b.

T. Evgeniou and M. Pontil. Regularized multi–task learning. In *Proc. KDD*, pages 109–117. ACM, 2004.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

C. Farabet, C. Couprie, L. Najman, and Y. LeCun. Learning hierarchical features for scene labeling. *IEEE Trans. PAMI*, 2013.

M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Stanford University, 2002.

L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proc. CVPR*, pages 2241–2248. IEEE, 2010a.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. PAMI*, 32(9):1627–1645, 2010b.

S. Fidler, R. Mottaghi, A. Yuille, and R. Urtasun. Bottom-up segmentation for top-down detection. In *Proc. CVPR*, pages 3294–3301. IEEE, 2013.

F. Fogel, R. Jenatton, F. Bach, and A. d'Aspremont. Convex relaxations for permutation problems. In *Proc. NIPS*, pages 1016–1024, 2013.

S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

R. Foygel and N. Srebro. Concentration-based guarantees for low-rank matrix reconstruction. In *Proc. COLT*, 2011.

R. Foygel, N. Srebro, and R. Salakhutdinov. Matrix reconstruction with the local max norm. In *Proc. NIPS*, pages 935–943, 2012.

Y. Freund, R. Schapire, and N. Abe. A short introduction to boosting. *Japanese Society For Artificial Intelligence*, 14(771-780):1612, 1999.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432, 2008.

J. Gallier. Notes on convex sets, polytopes, polyhedra, combinatorial topology, voronoi diagrams and delaunay triangulations. *arXiv preprint arXiv:0805.0292*, 2008.

P. Gehler and S. Nowozin. Infinite kernel learning. Technical report, Max Planck Institute For Biological Cybernetics, 2008.

P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *Proc. ICCV*, pages 221–228. IEEE, 2009.

A. Giannopoulos, V. D. Milman, and A. Tsolomitis. Asymptotic formulas for the diameter of sections of symmetric convex bodies. *Journal of Functional Analysis*, 223(1):86–108, 2005.

R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proc. CVPR*, 2014.

R. Girshick and J. Malik. Training deformable part models with decorrelated features. In *Proc. ICCV*, pages 3016–3023. IEEE, 2013.

T. Glasmachers. Universal consistency of multi-class support vector classification. In *Advances in Neural Information Processing Systems*, pages 739–747, 2010.

M. Gönen and E. Alpaydin. Localized multiple kernel learning. In *Proc. ICML*, pages 352–359. ACM, 2008.

M. Gönen and E. Alpaydın. Multiple kernel learning algorithms. *Journal of Machine Learning Research*, 12:2211–2268, 2011.

M. Gönen and E. Alpaydın. Localized algorithms for multiple kernel learning. *Pattern Recognition*, 46(3):795–807, 2013.

Y. Gordon. On milman's inequality and random subspaces which escape through a mesh in ? n. In *Geometric Aspects of Functional Analysis*. Springer, 1988.

E. Grave, G. R. Obozinski, and F. R. Bach. Trace lasso: a trace norm regularization for correlated designs. In *Proc. NIPS*, pages 2187–2195, 2011.

P. D. Grünwald. *The minimum description length principle*. MIT press, 2007.

A. Gupta, R. Nowak, and B. Recht. Sample complexity for 1-bit compressed sensing and sparse classification. In *ISIT*, pages 1553–1557, 2010.

P. Hall and J. Jin. Innovated higher criticism for detecting sparse signals in correlated noise. *Annals of Statistics*, 38(3):1686–1732, 2010.

T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, pages 1391–1415, 2004.

J. Haupt, R. Castro, and R. Nowak. Distilled sensing: Selective sampling for sparse signal recovery. In *Proc. AISTATS*, pages 216–223, 2009.

J. Haupt, R. Baraniuk, R. Castro, and R. Nowak. Sequentially designed compressed sensing. In *Statistical Signal Processing Workshop (SSP), 2012 IEEE*, pages 401–404. IEEE, 2012.

E. Hazan. Sparse approximate solutions to semidefinite programs. In *Proceedings of the 8th Latin American conference on Theoretical informatics*, pages 306–316. Springer-Verlag, 2008.

K. He, X. Zhang, S. Ren, and J. Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *arXiv preprint arXiv:1406.4729*, 2014.

A. Hero and B. Rajaratnam. Hub discovery in partial correlation graphs. *IEEE Trans. Information Theory*, 58(9):6064–6078, 2012.

G. E. Hinton and S. Osindero. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:2006, 2006.

D. Hoiem, Y. Chodpathumwan, and Q. Dai. Diagnosing error in object detectors. In *Proc. ECCV*, pages 340–353. Springer, 2012.

P. O. Hoyer. Non-negative Matrix Factorization with Sparseness Constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004a.

P. O. Hoyer. Non-negative sparse coding. In *Neural Networks for Signal Processing, 2002. Proceedings of the 2002 12th IEEE Workshop on*, pages 557–565. IEEE, 2002.

P. O. Hoyer. Non-negative matrix factorization with sparseness constraints. *Journal of Machine Learning Research*, 5:1457–1469, 2004b.

P. Hu, S. Janga, M. Babu, J. Díaz-Mejía, G. Butland, W. Yang, O. Pogoutse, X. Guo, S. Phanse, P. Wong, et al. Global functional atlas of escherichia coli encompassing previously uncharacterized proteins. *PLoS biology*, 7(4), 2009.

F. Iandola, M. Moskewicz, S. Karayev, R. Girshick, T. Darrell, and K. Keutzer. Densenet: Implementing efficient convnet descriptor pyramids. *arXiv preprint arXiv:1404.1869*, 2014.

Y. Ingster. Some problem of hypothesis testing leading to infinitely divisible distributions. *Mathematical Methods of Statistics*, 6:47–69, 1997.

M. Jaggi. Revisiting {Frank-Wolfe}: Projection-free sparse convex optimization. In *Proc. ICML*, pages 427–435, 2013.

A. Jalali and N. Srebro. Clustering using max-norm constrained optimization. In *Proc. ICML*, 2012.

Y. Jalali, A.and Chen, S. Sanghavi, and H. Xu. Clustering partially observed graphs via convex optimization. In *Proc. ICML*, 2011.

G. J. O. Jameson. *Summing and nuclear norms in Banach space theory*. Number 8. Cambridge University Press, 1987.

I. M. Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of Statistics*, 29(2):295–327, 2001.

I. M. Johnstone and A. Y. Lu. On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, 104(486), 2009.

I. A. Kakadiaris, G. Passalis, G. Toderici, M. N. Murtuza, Y. Lu, N. Karampatziakis, and T. Theoharis. Three-dimensional face recognition in the presence of facial expressions: An annotated deformable model approach. *IEEE Trans. PAMI*, 29(4):640–649, 2007.

M. H. Kamal and P. Vandergheynst. Joint low-rank and sparse light field modelling for dense multiview data compression. In *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, pages 3831–3835. Ieee, 2013.

N. E. Karoui. Operator norm consistent estimation of large-dimensional sparse covariance matrices. *Annals of Statistics*, pages 2717–2756, 2008.

J. Kim and H. Park. Sparse nonnegative matrix factorization for clustering. Technical report, Georgia Institute of Technology, 2008.

M. Kloft, U. Brefeld, S. Sonnenburg, P. Laskov, K.-R. Müller, and A. Zien. Efficient and accurate lp-norm multiple kernel learning. In *Proc. NIPS*, volume 22, pages 997–1005, 2009.

I. Kokkinos. Rapid deformable object detection using dual-tree branch-and-bound. In *Proc. NIPS*, pages 2681–2689, 2011.

I. Kokkinos. Shufflets: shared mid-level parts for fast object detection. In *Proc. ICCV*, pages 1393–1400. IEEE, 2013.

V. Koltchinskii, K. Lounici, and A. Tsybakov. Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics*, 2011a.

V. Koltchinskii, K. Lounici, A. B. Tsybakov, et al. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39(5):2302–2329, 2011b.

Y. Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proc. KDD*, pages 426–434. ACM, 2008.

A. Krizhevsky, I. Sutskever, G., and Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In *Proc. NIPS*, 2012.

B. Kulis. Metric learning: A survey. *Foundations and Trends® in Machine Learning*, 5(4): 287–364, 2013. ISSN 1935-8237. doi: 10.1561/2200000019. URL http://dx.doi.org/10.1561/2200000019.

G. Lanckriet, N. Cristianini, P. Bartlett, and L. E. Ghaoui. Learning the Kernel Matrix with Semi-Definite Programming. *Journal of Machine Learning Research*, 5:2004, 2002.

D. Lee, H. Seung, et al. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

T. Lee, A. Shraibman, and R. Spalek. A direct product theorem for discrepancy. In *Computational Complexity, 2008. CCC'08. 23rd Annual IEEE Conference on*, pages 71–80. IEEE, 2008.

M. Li and P. M. Vitányi. *An introduction to Kolmogorov complexity and its applications*. Springer, 2009.

D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

J. Liu, P. Musialski, P. Wonka, and J. Ye. Tensor completion for estimating missing values in visual data. *IEEE Trans. PAMI*, 35(1):208–220, 2013.

D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. ICCV*, volume 2, pages 1150–1157. Ieee, 1999.

X. Luo. High dimensional low rank and sparse covariance matrix estimation via convex minimization. *Arxiv preprint arXiv:1111.1133*, 2011.

L. W. Mackey. Deflation methods for sparse pca. In *Proc. NIPS*, pages 1017–1024, 2009.

S. Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

S. G. Mallat and Z. Zhang. Matching pursuits with time-frequency dictionaries. *Signal Processing, IEEE Transactions on*, 41(12):3397–3415, 1993.

C. L. Mallows. Some comments on c p. *Technometrics*, 15(4):661–675, 1973.

M. Malloy and R. Nowak. On the limits of sequential testing in high dimensions. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 1245–1249. IEEE, 2011a.

M. Malloy and R. Nowak. Sequential analysis in high-dimensional multiple testing and sparse recovery. In *Information Theory Proceedings (ISIT), 2011 IEEE International Symposium on*, pages 2661–2665. IEEE, 2011b.

V. Milman. Random subspaces of proportional dimension of finite dimensional normed spaces: approach through the isoperimetric inequality. In *Banach spaces*, pages 106–115. Springer, 1985.

V. Milman. Surprising geometric phenomena in high-dimensional convexity theory. In *European Congress of Mathematics*, pages 73–91. Springer, 1998.

M.-E. Nilsback and A. Zisserman. Automated flower classification over a large number of classes. In *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing*, pages 722–729. IEEE, 2008.

T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. PAMI*, 24(7):971–987, 2002.

I. Olkin and J. Pratt. A multivariate tchebycheff inequality. *The Annals of Mathematical Statistics*, 29(1):226–234, 1958.

C. S. Ong, R. C. Williamson, and A. J. Smola. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, pages 1043–1071, 2005.

S. Oymak, A. Jalali, M. Fazel, Y. C. Eldar, and B. Hassibi. Simultaneously structured models with application to sparse and low-rank matrices. *arXiv preprint arXiv:1212.3753*, 2012.

A. Pajor and N. Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional banach spaces. *Proceedings of the American Mathematical Society*, 97(4): 637–642, 1986.

N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 123–231, 2013.

F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *Proc. CVPR*, pages 3482–3489. IEEE, 2012.

Y. Plan and R. Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013a.

Y. Plan and R. Vershynin. Robust 1-bit compressed sensing and sparse logistic regression: A convex programming approach. *IEEE Trans. Information Theory*, 59(1):482–494, 2013b.

J. Platt. *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA, 1999. ISBN 0-262-19416-3.

H. Raguet, J. Fadili, and G. Peyré. A generalized forward-backward splitting. *SIAM Journal on Imaging Sciences*, 6(3):1199–1226, 2013.

A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9(11), 2008.

M. Ranzato, J. Susskind, V. Mnih, and G. Hinton. On deep generative models with applications to recognition. In *Proc. CVPR*, 2011.

N. S. Rao, B. Recht, and R. D. Nowak. Universal measurement bounds for structured sparse signal recovery. In *Proc. AISTATS*, pages 942–950, 2012.

E. Richard, N. Baskiotis, T. Evgeniou, and N. Vayatis. Link discovery using graph feature tracking. *Proc. NIPS*, 2010.

E. Richard, P.-A. Savalle, and N. Vayatis. Estimation of simultaneously sparse and low rank matrices. In *Proc. ICML*, 2012.

E. Richard, F. Bach, J.-P. Vert, et al. Intersecting singularities for multi-structured estimation. In *Proc. ICML*, 2013.

R. T. Rockafellar. *Convex analysis*. Number 28. Princeton university press, 1997.

B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. In *Proc. NIPS*, pages 2967–2975, 2013.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, pages 1012–1030, 2007.

P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun. Pedestrian detection with unsupervised multi-stage feature learning. In *Proc. CVPR*, 2013.

P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, and Y. LeCun. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *International Conference on Learning Representations (ICLR 2014)*. CBLS, April 2014.

F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons. Document clustering using nonnegative matrix factorization. *Information Processing & Management*, 42(2):373–386, 2006.

G. R. Shorack and J. A. Wellner. *Empirical processes with applications to statistics*, volume 59. SIAM, 2009.

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

S. Sra. Fast projections onto $\ell_{1,q}$-norm balls for grouped feature selection. In *Machine learning and knowledge discovery in databases*, pages 305–317. Springer, 2011.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *Learning Theory*, pages 599–764, 2005.

N. Srebro, J. Rennie, and T. Jaakkola. Maximum-margin matrix factorization. *Proc. NIPS*, 17:1329–1336, 2005.

N. Srebro. *Learning with matrix factorizations*. PhD thesis, Massachusetts Institute of Technology, 2004.

N. Srebro and R. Salakhutdinov. Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Proc. NIPS*, pages 2056–2064, 2010.

N. Srinivas, A. Krause, S. Kakade, and M. Seeger. Gaussian process optimization in the bandit setting: No regret and experimental design. In *Proc. ICML*, 2010.

I. Steinwart. Consistency of support vector machines and other regularized kernel classifiers. *IEEE Trans. Information Theory*, 51(1):128–142, 2005.

I. Steinwart and A. Christmann. *Support vector machines*. Springer, 2008.

I. Steinwart, D. Hush, and C. Scovel. Training svms without offset. *Journal of Machine Learning Research*, 12:141–202, 2011.

Y. Tang, R. Salakhutdinov, and G. Hinton. Deep lambertian networks. In *Proc. ICML*, 2012.

B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller. Link prediction in relational data. In *Proc. NIPS*, page None, 2003.

A. Tewari and P. L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996a.

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

R. Tibshirani, J. Friedman, and T. Hastie. *The Elements of Statistical Learning*. Springer, 2009.

N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *37th annual Allerton Conference on Communication, Control, and Computing*, 1999.

A. L. Traud, P. J. Mucha, and M. A. Porter. Social structure of facebook networks. *Physica A: Statistical Mechanics and its Applications*, 391(16):4165–4180, 2012.

A. B. Tsybakov. *Introduction to nonparametric estimation*. Springer, 2009.

J. R. Uijlings, K. E. van de Sande, T. Gevers, and A. W. Smeulders. Selective search for object recognition. *International journal of computer vision*, 104(2):154–171, 2013.

S. Vattikuti, J. J. Lee, C. C. Chang, S. D. Hsu, and C. C. Chow. Applying compressed sensing to genome-wide association studies. *GigaScience*, 3(1):10, 2014.

A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *Proc. ICCV*, pages 606–613. IEEE, 2009.

R. Vershynin. Lectures in geometric functional analysis. *Unpublished manuscript. Available at http://www-personal. umich. edu/˜ romanv/papers/GFA-book/GFA-book. pdf*, 2011.

P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. CVPR*, volume 1, pages I–511. IEEE, 2001.

U. Von Luxburg. A tutorial on spectral clustering. *Statistics and computing*, 17(4):395–416, 2007.

C. Vondrick, A. Khosla, T. Malisiewicz, and A. Torralba. Hoggles: Visualizing object detection features. In *Proc. ICCV*, pages 1–8. IEEE, 2013.

M. Wang and A. Tang. Conditions for a unique non-negative solution to an underdetermined system. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 301–307. IEEE, 2009.

W. Xu, X. Liu, and Y. Gong. Document clustering based on non-negative matrix factorization. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 267–273. ACM, 2003.

J. Yang, Y. Li, Y. Tian, L. Duan, and W. Gao. Group-sensitive multiple kernel learning for object categorization. In *Proc. ICCV*, pages 436–443. IEEE, 2009.

J. Yang, Y. Li, Y. Tian, L.-Y. Duan, and W. Gao. Per-sample multiple kernel approach for visual concept learning. *Journal on Image and Video Processing*, 2010:2, 2010.

J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 197–206. ACM, 2007.

L. Yang and R. Jin. Distance metric learning: a comprehensive survey. Technical report, Michigan State University, May 2006.

Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *Proc. CVPR*, pages 1385–1392. IEEE, 2011.

Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures of parts. *IEEE Trans. PAMI*, 35(12), 2013.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

C.-H. Zhang. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics*, pages 894–942, 2010.

X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, 2012a.

X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *Proc. CVPR*, pages 2879–2886. IEEE, 2012b.

H. Zou, T. Hastie, and R. Tibshirani. Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, pages 1–30, 2004.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005a.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005b.