High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning

Francis Bach

High-Dimensional Non-Linear Variable Selection through Hierarchical Kernel Learning

Francis Bach

INRIA - WILLOW Project-Team
Laboratoire d'Informatique de l'Ecole Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)
23, avenue d'Italie, 75214 Paris, France
francis.bach@inria.fr

September 4, 2009

Abstract

We consider the problem of high-dimensional non-linear variable selection for supervised learning. Our approach is based on performing linear selection among exponentially many appropriately defined positive definite kernels that characterize non-linear interactions between the original variables. To select efficiently from these many kernels, we use the natural hierarchical structure of the problem to extend the multiple kernel learning framework to kernels that can be embedded in a directed acyclic graph; we show that it is then possible to perform kernel selection through a graph-adapted sparsity-inducing norm, in polynomial time in the number of selected kernels. Moreover, we study the consistency of variable selection in high-dimensional settings, showing that under certain assumptions, our regularization framework allows a number of irrelevant variables which is exponential in the number of observations. Our simulations on synthetic datasets and datasets from the UCI repository show state-of-the-art predictive performance for non-linear regression problems.

1 Introduction

High-dimensional problems represent a recent and important topic in machine learning, statistics and signal processing. In such settings, some notion of sparsity is a fruitful way of avoiding overfitting, for example through variable or feature selection. This has led to many algorithmic and theoretical advances. In particular, regularization by sparsity-inducing norms such as the ℓ_1-norm has attracted a lot of interest in recent years. While early work has focused on efficient algorithms to solve the convex optimization problems, recent research has looked at the model selection properties and predictive performance of such methods, in the linear case (Zhao and Yu, 2006; Yuan and Lin, 2007; Zou, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009a) or within constrained non-linear settings such as the multiple kernel learning framework (Lanckriet et al., 2004b; Srebro and Ben-David, 2006; Bach, 2008a; Koltchinskii and Yuan, 2008; Ying and Campbell, 2009) or generalized additive models (Ravikumar et al., 2008; Lin and Zhang, 2006).

However, most of the recent work dealt with *linear high-dimensional* variable selection, while the focus of much of the earlier work in machine learning and statistics was on *non-linear low-dimensional* problems: indeed, in the last two decades, kernel methods have been a prolific theoretical and algorithmic machine learning framework. By using appropriate regularization by Hilbertian norms, representer theorems enable to consider large and potentially infinite-dimensional feature spaces while working within an implicit feature space no larger than the number of observations. This has led to numerous works on kernel design adapted to specific data types and generic kernel-based algorithms for many learning tasks (see, e.g., Schölkopf and Smola, 2002; Shawe-Taylor and Cristianini, 2004). However, while non-linearity is required in many domains such as computer vision or bioinformatics, most theoretical results related to non-parametric methods do not scale well with input dimensions. In this paper, our goal is to bridge the gap between linear and non-linear methods, by tackling *high-dimensional non-linear* problems.

The task of non-linear variable section is a hard problem with few approaches that have both good theoretical and algorithmic properties, in particular in high-dimensional settings. Among classical methods, some are implicitly or explicitly based on sparsity and model selection, such as boosting (Freund and Schapire, 1997), multivariate additive regression splines (Friedman, 1991), decision trees (Breiman et al., 1984), random forests (Breiman, 2001), Cosso (Lin and Zhang, 2006) or Gaussian process based methods (see, e.g., Rasmussen and Williams, 2006), while some others do not rely on sparsity, such as nearest neighbors or kernel methods (see, e.g., Devroye et al., 1996; Shawe-Taylor and Cristianini, 2004).

First attempts were made to combine non-linearity and sparsity-inducing norms by considering *generalized additive models*, where the predictor function is assumed to be a sparse linear combination of non-linear functions of each variable (Bach et al., 2004a; Bach, 2008a; Ravikumar et al., 2008). However, as shown in Section 5.3, higher orders of interactions are needed for universal consistency, i.e., to adapt to the potential high complexity of the interactions between the relevant variables; we need to potentially allow 2^p of them for p variables (for all possible subsets of the p variables). Theoretical results suggest that with appropriate assumptions, sparse methods such as greedy methods and methods based on the ℓ_1-norm would be able to deal correctly with 2^p features if p is of the order of the number of observations n (Wainwright, 2009; Candès and Wakin, 2008; Zhang, 2009b). However, in presence of more than a few dozen variables, in order to deal with that many features, or even to simply enumerate those, a certain form of factorization or recursivity is needed. In this paper, we propose to use a hierarchical structure based on directed acyclic graphs, which is natural in our context of non-linear variable selection.

We consider a positive definite kernel that can be expressed as a large sum of positive definite *basis* or *local kernels*. This exactly corresponds to the situation where a large feature space is the concatenation of smaller feature spaces, and we aim to do selection among these many kernels (or equivalently feature spaces), which may be done through multiple kernel learning (Bach et al., 2004a). One major difficulty however is that the number of these smaller kernels is usually exponential in the dimension of the input space and applying multiple kernel learning directly to this decomposition would be intractable. As shown in Section 3.2, for non-linear variable selection, we consider a sum of kernels which are indexed by the set of subsets of all considered variables, or more generally by $\{0, \dots, q\}^p$, for $q \geqslant 1$.

In order to perform selection efficiently, we make the extra assumption that these small kernels can be embedded in a *directed acyclic graph* (DAG). Following Zhao et al. (2009), we consider in Section 2 a specific combination of ℓ_2-norms that is adapted to the DAG, and that will restrict

the authorized sparsity patterns to certain configurations; in our specific kernel-based framework, we are able to use the DAG to design an optimization algorithm which has polynomial complexity in the number of selected kernels (Section 4). In simulations (Section 6), we focus on *directed grids*, where our framework allows to perform non-linear variable selection. We provide some experimental validation of our novel regularization framework; in particular, we compare it to the regular ℓ_2-regularization, greedy forward selection and non-kernel-based methods, and shows that it is always competitive and often leads to better performance, both on synthetic examples, and standard regression datasets from the UCI repository.

Finally, we extend in Section 5 some of the known consistency results of the Lasso and multiple kernel learning (Zhao and Yu, 2006; Bach, 2008a), and give a partial answer to the model selection capabilities of our regularization framework by giving necessary and sufficient conditions for model consistency. In particular, we show that our framework is adapted to estimating consistently only the *hull* of the relevant variables. Hence, by restricting the statistical power of our method, we gain computational efficiency. Moreover, we show that we can obtain scalings between the number of variables and the number of observations which are similar to the linear case (Wainwright, 2009; Candès and Wakin, 2008; Zhao and Yu, 2006; Yuan and Lin, 2007; Zou, 2006; Wainwright, 2009; Bickel et al., 2009; Zhang, 2009a): indeed, we show that our regularization framework may achieve non-linear variable selection consistency even with a number of variables p which is exponential in the number of observations n. Since we deal with 2^p kernels, we achieve consistency with a number of kernels which is *doubly* exponential in n. Moreover, for general directed acyclic graphs, we show that the total number of vertices may grow unbounded as long as the maximal out-degree (number of children) in the DAG is less than exponential in the number of observations.

This paper extends previous work (Bach, 2008b), by providing more background on multiple kernel learning, detailing all proofs, providing new consistency results in high dimension, and comparing our non-linear predictors with non-kernel-based methods.

Notation. Throughout the paper we consider Hilbertian norms $\|f\|$ for elements f of Hilbert spaces, where the specific Hilbert space can always be inferred from the context (unless otherwise stated). For rectangular matrices A, we denote by $\|A\|_{\mathrm{op}}$ its largest singular value. We denote by $\lambda_{\max}(Q)$ and $\lambda_{\min}(Q)$ the largest and smallest eigenvalue of a symmetric matrix Q. These are naturally extended to compact self-adjoint operators (Brezis, 1980; Conway, 1997).

Moreover, given a vector v in the product space $\mathcal{F}_1 \times \cdots \times \mathcal{F}_p$ and a subset I of $\{1, \ldots, p\}$, v_I denotes the vector in $(\mathcal{F}_i)_{i \in I}$ of elements of v indexed by I. Similarly, for a matrix A defined with $p \times p$ blocks adapted to $\mathcal{F}_1, \ldots, \mathcal{F}_p$, A_{IJ} denotes the submatrix of A composed of blocks of A whose rows are in I and columns are in J. Moreover, $|J|$ denotes the cardinal of the set J and $|\mathcal{F}|$ denotes the dimension of the Hilbert space $\mathcal{F}$. We denote by 1_n the n-dimensional vector of ones. We denote by $(a)_+ = \max\{0, a\}$ the positive part of a real number a. Besides, given matrices $A_1, \ldots, A_n$, and a subset I of $\{1, \ldots, n\}$, $\mathrm{Diag}(A)_I$ denotes the block-diagonal matrix composed of the blocks indexed by I. Finally, we let denote $\mathbb{P}$ and $\mathbb{E}$ general probability measures and expectations.

	Loss $\varphi_i(u_i)$	Fenchel conjugate $\psi_i(\beta_i)$
Least-squares regression	$\frac{1}{2}(y_i - u_i)^2$	$\frac{1}{2}\beta_i^2 + \beta_i y_i$
1-norm support vector regression (SVR)	$(\lvert y_i - u_i\rvert - \varepsilon)_+$	$\beta_i y_i + \lvert\beta_i\rvert\varepsilon$ if $\lvert\beta\rvert \leqslant 1$ $+\infty$ otherwise
2-norm support vector regression (SVR)	$\frac{1}{2}(\lvert y_i - u_i\rvert - \varepsilon)_+^2$	$\frac{1}{2}\beta_i^2 + \beta_i y_i + \lvert\beta_i\rvert\varepsilon$
Hüber regression	$\frac{1}{2}(y_i - u_i)^2$ if $\lvert y_i - u_i\rvert \leqslant \varepsilon$ $\varepsilon\lvert y_i - u_i\rvert - \frac{\varepsilon^2}{2}$ otherwise	$\frac{1}{2}\beta_i^2 + \beta_i y_i$ if $\lvert\beta_i\rvert \leqslant \varepsilon$ $+\infty$ otherwise
Logistic regression	$\log(1 + \exp(-y_i u_i))$	$(1+\beta_i y_i)\log(1+\beta_i y_i) - \beta_i y_i \log(-\beta_i y_i)$ if $\beta_i y_i \in [-1, 0]$, $+\infty$ otherwise
1-norm support vector machine (SVM)	$\max(0, 1 - y_i u_i)$	$y_i\beta_i$ if $\beta_i y_i \in [-1, 0]$ $+\infty$ otherwise
2-norm support vector machine (SVM)	$\frac{1}{2}\max(0, 1 - y_i u_i)^2$	$\frac{1}{2}\beta_i^2 + \beta_i y_i$ if $\beta_i y_i \leqslant 0$ $+\infty$ otherwise

Table 1: Loss functions with corresponding Fenchel conjugates, for regression (first three losses, $y_i \in \mathbb{R}$) and binary classification (last three losses, $y_i \in \{-1, 1\}$.

2 Review of Multiple Kernel Learning

We consider the problem a predicting a *response* $Y \in \mathbb{R}$ from a variable $X \in \mathcal{X}$, where $\mathcal{X}$ may be any set of inputs, referred to as the *input space*. In this section, we review the multiple kernel learning framework our paper relies on.

2.1 Loss Functions

We assume that we are given n observations of the couple (X, Y), i.e., $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ for $i = 1, \ldots, n$. We define the *empirical risk* of a function f from $\mathcal{X}$ to $\mathbb{R}$ as

$$\frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)),$$

where $\ell : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}^+$ is a *loss function*. We only assume that ℓ is convex with respect to the second parameter (but not necessarily differentiable).

Following Bach et al. (2004b) and Sonnenburg et al. (2006), in order to derive optimality conditions for all losses, we need to introduce Fenchel conjugates (see examples in Table 1 and Figure 1). Let $\psi_i : \mathbb{R} \mapsto \mathbb{R}$, be the Fenchel conjugate (Boyd and Vandenberghe, 2003) of the convex function $\varphi_i : u_i \mapsto \ell(y_i, u_i)$, defined as

$$\psi_i(\beta_i) = \max_{u_i \in \mathbb{R}} \; u_i\beta_i - \varphi_i(u_i) = \max_{u_i \in \mathbb{R}} \; u_i\beta_i - \ell(y_i, u_i).$$

The function ψ_i is always convex and, because we have assumed that φ_i is convex, we can represent φ_i as the Fenchel conjugate of ψ_i, i.e., for all $u_i \in \mathbb{R}$,

$$\ell(y_i, u_i) = \varphi_i(u_i) = \max_{\beta_i \in \mathbb{R}} \; u_i\beta_i - \psi_i(\beta_i).$$

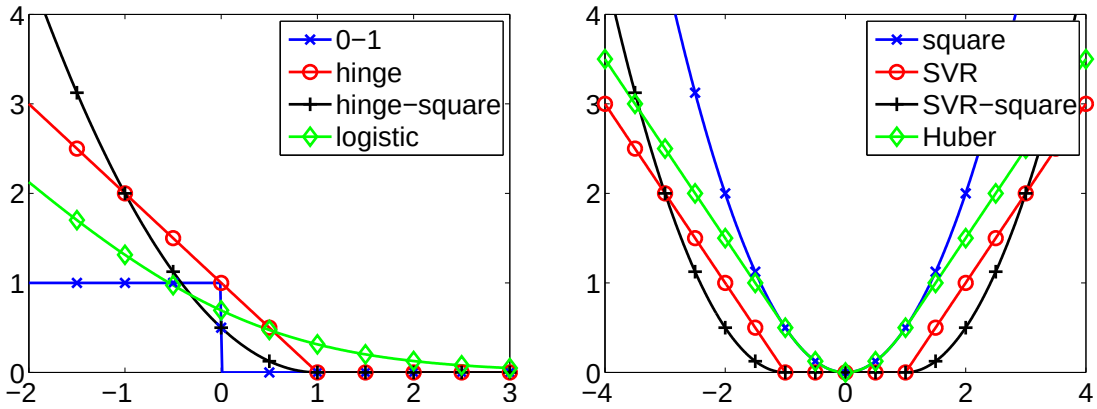

Figure 1: (Left) Losses for binary classification (plotted with $y_i = 1$). (Right) Losses for regression (plotted with $y_i = 0$).

Moreover, in order to include an unregularized constant term, we will need to be able to solve with respect to $b \in \mathbb{R}$ the following optimization problem:

$$\min_{b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \varphi_i(u_i + b). \tag{1}$$

For $u \in \mathbb{R}^n$, we let denote by $b^*(u)$ any solution of Eq. (1). It can either be obtained in closed form (least-squares regression), using Newton-Raphson (logistic regression), or by ordering the values $u_i \in \mathbb{R}$, $i = 1, \dots, n$ (all other piecewise quadratic losses). In Section 4, we study in details losses for which the Fenchel conjugate ψ_i is strictly convex, such as for logistic regression, 2-norm SVM, 2-norm SVR and least-squares regression.

2.2 Single Kernel Learning Problem

In this section, we assume that we are given a positive definite kernel $k(x, x')$ on $\mathcal{X}$. We can then define a reproducing kernel Hilbert space (RKHS) as the completion of the linear span of functions $x \mapsto k(x, x')$ for $x' \in \mathcal{X}$ (Berlinet and Thomas-Agnan, 2003). We can define the *feature map* $\Phi : \mathcal{X} \mapsto \mathcal{F}$ such that for all $x \in \mathcal{X}$, $f(x) = \langle f, \Phi(x) \rangle$ and for all $x, x' \in \mathcal{X}$, $\Phi(x)(x') = k(x, x')$; we denote by $\|f\|$ the norm of the function $f \in \mathcal{F}$. We consider the single kernel learning problem:

$$\min_{f \in \mathcal{F},\, b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, f(x_i) + b\right) + \frac{\lambda}{2} \|f\|^2. \tag{2}$$

The following proposition gives its dual, providing a convex instance of the representer theorem (see, e.g. Shawe-Taylor and Cristianini, 2004; Schölkopf and Smola, 2002, and proof in Appendix A.2):

Proposition 1 (Dual problem for single kernel learning problem) *The dual of the optimization problem in Eq. (2) is*

$$\max_{\alpha \in \mathbb{R}^n,\, 1_n^\top \alpha = 0} -\frac{1}{n} \sum_{i=1}^{n} \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2} \alpha^\top K \alpha, \tag{3}$$

where $K \in \mathbb{R}^{n\times n}$ is the kernel matrix defined as $K_{ij} = k(x_i, x_j)$. The unique primal solution f can be found from an optimal α as $f = \sum_{i=1}^n \alpha_i \Phi(x_i)$, and $b = b^*(K\alpha)$.

Note that if the Fenchel conjugate is strictly convex or if the kernel matrix is invertible, then the dual solution α is also unique. In Eq. (3), the kernel matrix K may be replaced by its *centered* version

$$\widetilde{K} = \Big(\mathrm{I} - \frac{1}{n} 1_n 1_n^\top\Big) K \Big(\mathrm{I} - \frac{1}{n} 1_n 1_n^\top\Big),$$

defined as the kernel matrix of the centered observed features (see, e.g. Shawe-Taylor and Cristianini, 2004; Schölkopf and Smola, 2002). Indeed, we have $\alpha^\top \widetilde{K}\alpha = \alpha^\top K \alpha$ in Eq. (3); however, in the definition of $b = b^*(K\alpha)$, K cannot be replaced by $\widetilde{K}$.

Finally, the duality gap obtained from a vector $\alpha \in \mathbb{R}^n$ such that $1_n^\top \alpha = 0$, and the associated primal candidates from Proposition 1 is equal to

$$\mathrm{gap}_{\mathrm{kernel}}\,(K, \alpha) = \frac{1}{n}\sum_{i=1}^n \varphi_i\left[(K\alpha)_i + b^*(K\alpha)\right] + \lambda \alpha^\top \widetilde{K}\alpha + \frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i). \qquad (4)$$

2.3 Sparse Learning with Multiple Kernels

We now assume that we are given p different reproducing kernel Hilbert spaces $\mathcal{F}_j$ on $\mathcal{X}$, associated with positive definite kernels $k_j : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, $j = 1, \ldots, p$, and associated feature maps $\Phi_j : \mathcal{X} \to \mathcal{F}_j$. We consider generalized additive models (Hastie and Tibshirani, 1990), i.e., predictors parameterized by $f = (f_1, \ldots, f_p) \in \mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_p$ of the form

$$f(x) + b = \sum_{j=1}^p f_j(x) + b = \sum_{j=1}^p \langle f_j, \Phi_j(x)\rangle + b,$$

where each $f_j \in \mathcal{F}_j$ and $b \in \mathbb{R}$ is a constant term. We let denote $\|f\|$ the Hilbertian norm of $f \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_p$, defined as $\|f\|^2 = \sum_{j=1}^p \|f_j\|^2$.

We consider regularizing by the sum of the Hilbertian norms, $\sum_{j=1}^p \|f_j\|$ (which is not itself a Hilbertian norm), with the intuition that this norm will push some of the functions f_j towards zero, and thus provide data-dependent selection of the feature spaces $\mathcal{F}_j$, $j = 1, \ldots, p$, and hence selection of the kernels k_j, $j = 1, \ldots, p$. We thus consider the following optimization problem:

$$\min_{f_1\in\mathcal{F}_1,\ \ldots,f_p\in\mathcal{F}_p,\ b\in\mathbb{R}} \ \frac{1}{n}\sum_{i=1}^n \ell\Big(y_i, \sum_{j=1}^p f_j(x_i) + b\Big) + \frac{\lambda}{2}\Big(\sum_{j=1}^p \|f_j\|\text{‘} \qquad (5)$$

Note that using the squared sum of norms does not change the regularization properties: for all solutions of the problem regularized by $\sum_{j=1}^p \|f_j\|$, there corresponds a solution of the problem in Eq. (5) with a different regularization parameter, and vice-versa (see, e.g., Borwein and Lewis, 2000, Section 3.2). The previous formulation encompasses a variety of situations, depending on how we set up the input spaces $\mathcal{X}_1, \ldots, \mathcal{X}_p$:

- **Regular ℓ_1-norm and group ℓ_1-norm regularization**: if each $\mathcal{X}_j$ is the space of real numbers, then we exactly get back penalization by the ℓ_1-norm, and for the square loss, the Lasso (Tibshirani, 1996); if we consider finite dimensional vector spaces, we get back the block ℓ_1-norm formulation and the group Lasso for the square loss (Yuan and Lin, 2006). Our general Hilbert space formulation can thus be seen as a "non-parametric group Lasso".

- **"Multiple input space, multiple feature spaces"**: In this section, we assume that we have a single input space $\mathcal{X}$ and multiple feature spaces $\mathcal{F}_1, \dots, \mathcal{F}_p$ defined on the same input space. We could also consider that we have p different input spaces $\mathcal{X}_j$ and one feature space $\mathcal{F}_j$ per $\mathcal{X}_j$, $j = 1, \dots, p$, a situation common in generalized additive models. We can go from the "single input space, multiple feature spaces" view to the "multiple input space/feature space pairs" view by considering p identical copies $\mathcal{X}_1, \dots, \mathcal{X}_p$ or $\mathcal{X}$, while we can go in the other direction using projections from $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$.

The sparsity-inducing norm formulation defined in Eq. (5) can be seen from several points of views and this has led to interesting algorithmic and theoretical developments, which we review in the next sections. In this paper, we will build on the approach of Section 2.4, but all results could be derived through the approach presented in Section 2.5 and Section 2.6.

2.4 Learning convex combinations of kernels

Pontil and Micchelli (2005) and Rakotomamonjy et al. (2008) show that

$$\Big(\sum_{j=1}^p \|f_j\| \Big)^2 = \min_{\zeta \in \mathbb{R}_+^p,\ 1_p^\top \zeta = 1} \sum_{j=1}^p \frac{\|f_j\|^2}{\zeta_j},$$

where the minimum is attained at $\zeta_j = \|f_j\| / \sum_{k=1}^p \|f_k\|$. This variational formulation of the squared sum of norms allows to find an equivalent problem to Eq. (5), namely:

$$\min_{\zeta \in \mathbb{R}_+^p,\ 1_p^\top \zeta = 1} \ \min_{f_1 \in \mathcal{F}_1, \dots, f_p \in \mathcal{F}_p,\ b \in \mathbb{R}} \ \frac{1}{n} \sum_{i=1}^n \ell\Big(y_i, \sum_{j=1}^p f_j(x_i) + b \Big) + \frac{\lambda}{2} \sum_{j=1}^p \frac{\|f_j\|^2}{\zeta_j}. \tag{6}$$

Given $\zeta \in \mathbb{R}_+^p$ such that $1_p^\top \zeta = 1$, using the change of variable $\tilde{f}_j = f_j \zeta_j^{-1/2}$ and $\tilde{\Phi}_j(x) = \zeta_j^{1/2} \Phi_j(x)$, $j = 1, \dots, p$, the problem in Eq. (6) is equivalent to:

$$\min_{\zeta \in \mathbb{R}_+^p,\ 1_p^\top \zeta = 1} \ \min_{\tilde{f} \in \mathcal{F},\ b \in \mathbb{R}} \ \frac{1}{n} \sum_{i=1}^n \ell\big(y_i, \langle \tilde{f}, \tilde{\Phi}(x_i) \rangle + b \big) + \frac{\lambda}{2} \|\tilde{f}\|^2,$$

with respect to $\tilde{f}$. Thus $\tilde{f}$ is the solution of the single kernel learning problem with kernel

$$k(\zeta)(x, x') = \langle \tilde{\Phi}(x), \tilde{\Phi}(x') \rangle = \sum_{j=1}^p \langle \zeta_j^{1/2} \Phi_j(x), \zeta_j^{1/2} \Phi_j(x') \rangle = \sum_{j=1}^p \zeta_j k_j(x, x').$$

This shows that the non-parametric group Lasso formulation amounts in fact to learning implicitly a weighted combination of kernels (Bach et al., 2004a; Rakotomamonjy et al., 2008). Moreover, the optimal functions f_j can then be computed as $f_j(\cdot) = \zeta_j \sum_{i=1}^n \alpha_i k_j(\cdot, x_i)$, where the vector $\alpha \in \mathbb{R}^n$ is *common* to all feature spaces $\mathcal{F}_j$, $j = 1, \dots, p$.

2.5 Conic convex duality

One can also consider the convex optimization problem in Eq. (5) and derive the convex dual using conic programming (Lobo et al., 1998; Bach et al., 2004a; Bach, 2008a):

$$\max_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0} \left\{ -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2}\max_{j\in\{1,\ldots,p\}} \alpha^\top \widetilde{K}_j\alpha \right\}, \tag{7}$$

where $\widetilde{K}_j$ is the centered kernel matrix associated with the j-th kernel. From the optimality conditions for second order cones, one can also get that there exists positive weights ζ that sum to one, such that $f_j(\cdot) = \zeta_j \sum_{i=1}^n \alpha_i k_j(\cdot, x_i)$ (see Bach et al., 2004a, for details). Thus, both the kernel weights ζ and the solution α of the correspond learning problem can be derived from the solution of a single convex optimization problem based on second-order cones. Note that this formulation may be actually solved for small n with general-purpose toolboxes for second-order cone programming, although QCQP approaches may be used as well (Lanckriet et al., 2004a).

2.6 Kernel Learning with Semi-definite Programming

There is another way of seeing the same problem. Indeed, the dual problem in Eq. (7) may be rewritten as follows:

$$\max_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0}\ \min_{\zeta\in\mathbb{R}_+^p,\ 1_p^\top\zeta=1} \left\{ -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2}\alpha^\top\left(\sum_{j=1}^p \zeta_j\widetilde{K}_j\right)\alpha \right\}, \tag{8}$$

and by convex duality (Boyd and Vandenberghe, 2003; Rockafellar, 1970) as:

$$\min_{\zeta\in\mathbb{R}_+^p,\ 1_p^\top\zeta=1}\ \max_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0} \left\{ -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2}\alpha^\top\left(\sum_{j=1}^p \zeta_j\widetilde{K}_j\right)\alpha \right\}. \tag{9}$$

If we denote $G(K) = \max_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0} \left\{ -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\alpha_i) - \frac{\lambda}{2}\alpha^\top\widetilde{K}\alpha \right\}$, the optimal value of the single kernel learning problem in Eq. (2) with loss ℓ and kernel matrix K (and centered kernel matrix $\widetilde{K}$), then the multiple kernel learning problem is equivalent to minimizing $G(K)$ over convex combinations of the p kernel matrices associated with all p kernels, i.e., equivalent to minimizing $B(\zeta) = G(\sum_{j=1}^p \zeta_j K_j)$.

This function $G(K)$, introduced by several authors in slightly different contexts (Lanckriet et al., 2004b; Pontil and Micchelli, 2005; Ong et al., 2005), leads to a more general kernel learning framework where one can learn more than simply convex combinations of kernels—in fact, any kernel matrix which is positive semi-definite. In terms of theoretical analysis, results from general kernel classes may be brought to bear (Lanckriet et al., 2004b; Srebro and Ben-David, 2006; Ying and Campbell, 2009); however, the special case of convex combination allows the sparsity interpretation and some additional theoretical analysis (Bach, 2008a; Koltchinskii and Yuan, 2008). The practical and theoretical advantages of allowing more general potentially non convex combinations (not necessarily with positive coefficients) of kernels is still an open problem and subject of ongoing work (see, e.g., Varma and Babu, 2009, and references therein).

Note that regularizing in Eq. (5) by the sum of squared norms $\sum_{j=1}^p \|f_j\|^2$ (instead of the squared sum of norms), is equivalent to considering the sum of kernels matrices, i.e., $K = \sum_{j=1}^p K_j$.

Moreover, if all kernel matrices have rank one, then the kernel learning problem is equivalent to an ℓ_1-norm problem, for which dedicated algorithms are usually much more efficient (see, e.g., Efron et al., 2004; Wu and Lange, 2008).

2.7 Algorithms

The multiple facets of the multiple kernel learning problem have led to multiple algorithms. The first ones were based on the minimization of $B(\zeta) = G(\sum_{j=1}^{p} \zeta_j K_j)$ through general-purpose toolboxes for semidefinite programming (Lanckriet et al., 2004b; Ong et al., 2005). While this allows to get a solution with high precision, it is not scalable to medium and large-scale problems. Later, approaches based on conic duality and smoothing were derived (Bach et al., 2004a,b). They were based on existing efficient techniques for the support vector machine (SVM) or potentially other supervised learning problems, namely sequential minimal optimization (Platt, 1998). Although they are by design scalable, they require to recode existing learning algorithms and do not reuse pre-existing implementations. The latest formulations based on the direct minimization of a cost function that depends directly on ζ allow to reuse existing code (Sonnenburg et al., 2006; Rakotomamonjy et al., 2008) and may thus benefit from the intensive optimizations and tweaks already carried through. Finally, active set methods have been recently considered for finite groups (Roth and Fischer, 2008; Obozinski et al., 2009), an approach we extend to hierarchical kernel learning in Section 4.4.

3 Hierarchical Kernel Learning (HKL)

We now extend the multiple kernel learning framework to kernels which are indexed by vertices in a directed acyclic graph. We first describe examples of such graph-structured positive definite kernels from Section 3.1 to Section 3.4, and defined the graph-adapted norm in Section 3.5.

3.1 Graph-Structured Positive Definite Kernels

We assume that we are given a *positive definite kernel* $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, and that this kernel can be expressed as the sum, over an index set V, of basis kernels k_v, $v \in V$, i.e., for all $x, x' \in \mathcal{X}$:

$$k(x, x') = \sum_{v \in V} k_v(x, x').$$

For each $v \in V$, we denote by $\mathcal{F}_v$ and Φ_v the feature space and feature map of k_v, i.e., for all $x, x' \in \mathcal{X}$, $k_v(x, x') = \langle \Phi_v(x), \Phi_v(x') \rangle$.

Our sum assumption corresponds to a situation where the feature map $\Phi(x)$ and feature space $\mathcal{F}$ for k are the *concatenations* of the feature maps $\Phi_v(x)$ and feature spaces $\mathcal{F}_v$ for each kernel k_v, i.e., $\mathcal{F} = \prod_{v \in V} \mathcal{F}_v$ and $\Phi(x) = (\Phi_v(x))_{v \in V}$. Thus, looking for a certain $f \in \mathcal{F}$ and a predictor function $f(x) = \langle f, \Phi(x) \rangle$ is equivalent to looking jointly for $f_v \in \mathcal{F}_v$, for all $v \in V$, and

$$f(x) = \langle f, \Phi(x) \rangle = \sum_{v \in V} \langle f_v, \Phi_v(x) \rangle.$$

As mentioned earlier, we make the assumption that the set V can be embedded into a *directed acyclic graph*[1]. Directed acyclic graphs (referred to as DAGs) allow to naturally define the notions

[1]Throughout this paper, for simplicity, we use the same notation to refer to the graph and its set of vertices.

of *parents*, *children*, *descendants* and *ancestors* (Diestel, 2005). Given a node $w \in V$, we denote by $\mathrm{A}(w) \subset V$ the set of its ancestors, and by $\mathrm{D}(w) \subset V$, the set of its descendants. We use the convention that any w is a descendant and an ancestor of itself, i.e., $w \in \mathrm{A}(w)$ and $w \in \mathrm{D}(w)$. Moreover, for $W \subset V$, we let denote $\mathrm{sources}(W)$ the set of *sources* (or *roots*) of the graph V restricted to W, that is, nodes in W with no parents belonging to W.

Moreover, given a subset of nodes $W \subset V$, we can define the *hull* of W as the union of all ancestors of $w \in W$, i.e.,

$$\mathrm{hull}(W) = \bigcup_{w \in W} A(w).$$

Given a set W, we define the set of *extreme points* (or *sinks*) of W as the smallest subset $T \subset W$ such that $\mathrm{hull}(T) = \mathrm{hull}(W)$; it is always well defined, as (see Figure 2 for examples of these notions):

$$\mathrm{sinks}(W) = \bigcap_{T \subset V,\ \mathrm{hull}(T) = \mathrm{hull}(W)} T.$$

The goal of this paper is to perform kernel selection among the kernels k_v, $v \in V$. We essentially use the graph to limit the search to specific subsets of V. Namely, instead of considering all possible subsets of active (relevant) vertices, we will consider active sets of vertices which are equal to their hulls, i.e., subsets that contain the ancestors of all their elements, thus limiting the search space (see Section 3.5).

3.2 Decomposition of Usual Kernels in Directed Grids

In this paper, we primarily focus on kernels that can be expressed as "products of sums", and on the associated p-dimensional directed grids, while noting that our framework is applicable to many other kernels (see, e.g., Figure 4). Namely, we assume that the input space $\mathcal{X}$ factorizes into p components $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_p$ and that we are given p sequences of length $q+1$ of kernels $k_{ij}(x_i, x_i')$, $i \in \{1, \dots, p\}$, $j \in \{0, \dots, q\}$, such that (note the implicit different conventions for indices in k_i and k_{ij}):

$$k(x, x') = \prod_{i=1}^{p} k_i(x_i, x_i') = \prod_{i=1}^{p} \left(\sum_{j=0}^{q} k_{ij}(x_i, x_i') \right) = \sum_{j_1, \dots, j_p = 0}^{q} \prod_{i=1}^{p} k_{ij_i}(x_i, x_i'). \tag{10}$$

Note that in this section and the next section, x_i refers to the i-th component of the tuple $x = (x_1, \dots, x_p)$ (while in the rest of the paper, x_i is the i-th observation, which is itself a tuple). We thus have a sum of $(q+1)^p$ kernels, that can be computed efficiently as a product of p sums of $q+1$ kernels. A natural DAG on $V = \{0, \dots, q\}^p$ is defined by connecting each $(j_1, \dots, j_p)$ respectively to $(j_1+1, j_2, \dots, j_p)$, $\dots$, $(j_1, \dots, j_{p-1}, j_p+1)$ as long as $j_1 < q, \dots, j_p < q$, respectively . As shown in Section 3.5, this DAG (which has a single source) will correspond to the constraint of selecting a given product of kernels only after all the subproducts are selected. Those DAGs are especially suited to non-linear variable selection, in particular with the polynomial, Gaussian and spline kernels. In this context, products of kernels correspond to interactions between certain variables, and our DAG constraint implies that *we select an interaction only after all sub-interactions were already selected*, a constraint that is similar to the one used in multivariate additive splines (Friedman, 1991).

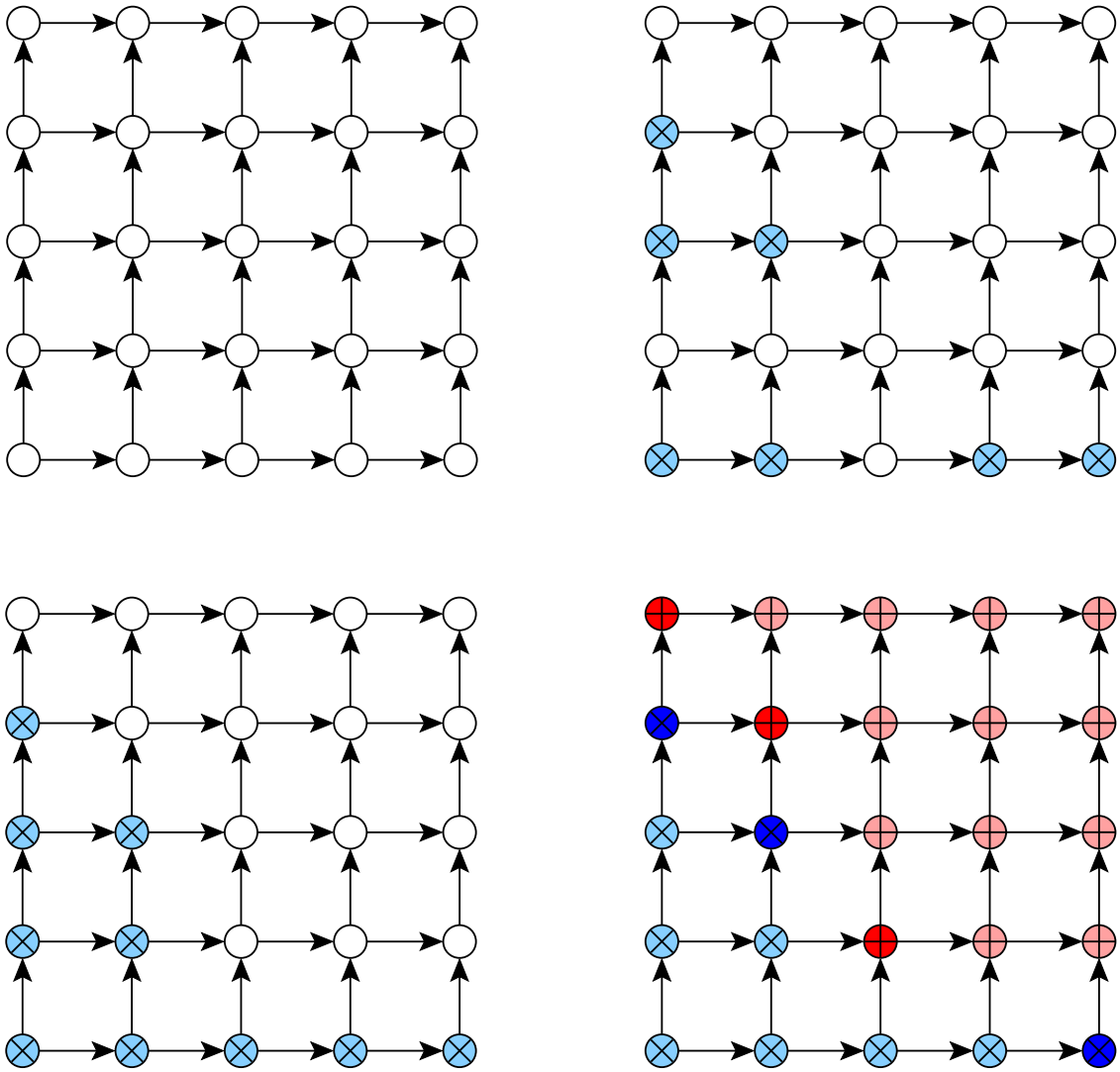

Figure 2: Examples of directed acyclic graphs (DAGs) and associated notions: (top left) 2D-grid (number of input variables $p = 2$, maximal order in each dimension $q = 4$); (top right) example of sparsity pattern which is not equal to its hull ($\times$ in light blue) and (bottom left) its hull ($\times$ in light blue); (bottom right) dark blue points ($\times$) are extreme points of the set of all active points (blue $\times$); dark red points ($+$) are the sources of the complement of the hull (set of all red $+$). Best seen in color.

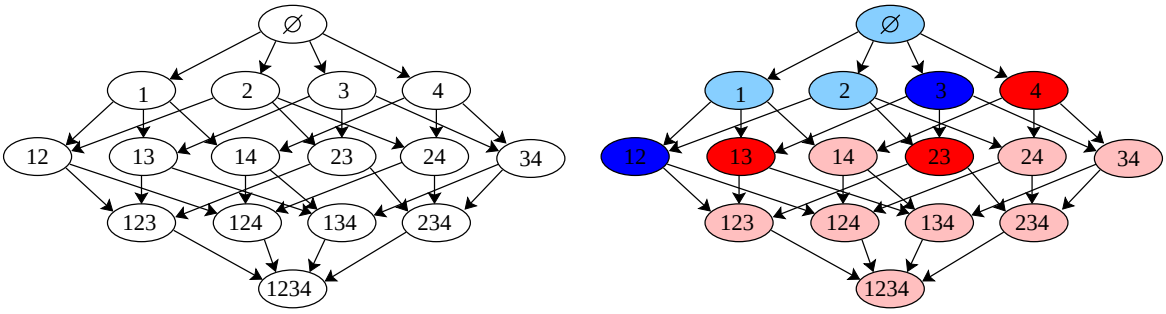

Figure 3: Directed acyclic graph of subsets of size 4: (left) DAG of subsets ($p = 4$, $q = 1$); (right) example of sparsity pattern (light and dark blue), dark blue points are extreme points of the set of all active points; dark red points are the sources of the set of all red points. Best seen in color.

Polynomial kernels. We consider $\mathcal{X}_i = \mathbb{R}$, $k_i(x_i, x_i') = (1 + x_i x_i')^q$ and for all $j \in \{0, \dots, q\}$, $k_{ij}(x_i, x_i') = \binom{q}{j}(x_i x_i')^j$; the full kernel is then equal to

$$k(x, x') = \prod_{i=1}^{p}(1 + x_i x_i')^q = \sum_{j_1, \dots, j_p = 0}^{q} \prod_{i=1}^{p} \binom{q}{j_i}(x_i x_i')^{j_i}.$$

Note that this is not exactly the usual polynomial kernel $(1 + x^\top x')^q$ (whose feature space is the space of multivariate polynomials of *total* degree less than q), since our kernel considers polynomials of *maximal* degree q.

Gaussian kernels (Gauss-Hermite decomposition). We also consider $\mathcal{X}_i = \mathbb{R}$, and the Gaussian-RBF kernel $e^{-b(x_i - x_i')^2}$ with $b > 0$. The following decomposition is the eigendecomposition of the non centered covariance operator corresponding to a normal distribution with variance $1/4a$ (see, e.g., Williams and Seeger, 2000; Bach, 2008a):

$$e^{-b(x_i - x_i')^2} = \left(1 - \frac{b^2}{A^2}\right)^{-1/2} \sum_{j=0}^{\infty} \frac{(b/A)^j}{2^j j!} e^{-\frac{b}{A}(a+c)x_i^2} H_j(\sqrt{2c}x_i) e^{-\frac{b}{A}(a+c)(x_i')^2} H_j(\sqrt{2c}x_i'), \quad (11)$$

where $c^2 = a^2 + 2ab$, $A = a + b + c$, and H_j is the j-th Hermite polynomial (Szegö, 1981). By appropriately truncating the sum, i.e., by considering that the first q basis kernels are obtained from the first q Hermite polynomials, and the $(q + 1)$-th kernel is summing over all other kernels, we obtain a decomposition of a uni-dimensional Gaussian kernel into $q + 1$ components (the first q of them are one-dimensional, the last one is infinite-dimensional, but can be computed by differencing). The decomposition ends up being close to a polynomial kernel of infinite degree, modulated by an exponential (Shawe-Taylor and Cristianini, 2004). One may also use an *adaptive* decomposition using kernel PCA (see, e.g. Shawe-Taylor and Cristianini, 2004; Schölkopf and Smola, 2002), which is equivalent to using the eigenvectors of the empirical covariance operator associated with the data (and not the population one associated with the Gaussian distribution with same variance). In prior work (Bach, 2008b), we tried both with no significant differences.

All-subset Gaussian kernels. When $q = 1$, the directed grid is isomorphic to the power set (i.e., the set of subsets, see Figure 3) with the DAG defined as the Hasse diagram of the partially ordered

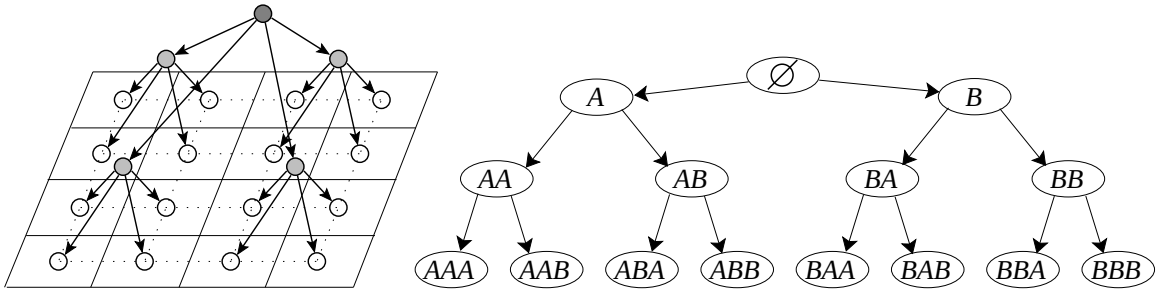

Figure 4: Additional examples of discrete structures. Left: pyramid over an image; a region is selected only after all larger regions that contains it are selected. Right: set of substrings of size 3 from the alphabet $\{A, B\}$; in bioinformatics (Schölkopf et al., 2004) and text processing (Lodhi et al., 2002), occurence of certain potentially long strings is an important feature and considering the structure may help selecting among the many possible strings.

set of all subsets (Cameron, 1994). In this setting, we can decompose the all-subset Gaussian kernel (see, e.g., Shawe-Taylor and Cristianini, 2004) as:

$$\prod_{i=1}^{p}(1+\alpha e^{-b(x_i-x_i')^2}) = \sum_{J\subset\{1,\dots,p\}} \prod_{i\in J} \alpha e^{-b(x_i-x_i')^2} = \sum_{J\subset\{1,\dots,p\}} \alpha^{|J|} e^{-b\|x_J-x_J'\|^2},$$

and our framework will select the relevant subsets for the Gaussian kernels, with the DAG presented in Figure 3. A similar decomposition is considered by Lin and Zhang (2006), but only on a subset of the power set. Note that the DAG of subsets is different from the "kernel graphs" introduced for the same type of kernel by Shawe-Taylor and Cristianini (2004) for expliciting the computation of polynomial kernels and ANOVA kernels.

Kernels on structured data. Although we mainly focus on directed grids in this paper, many kernels on structured data can also be naturally decomposed through a hierarchy (see Figure 4), such as the pyramid match kernel and related kernels (Grauman and Darrell, 2007; Cuturi and Fukumizu, 2006), string kernels or graph kernels (see, e.g., Shawe-Taylor and Cristianini, 2004). The main advantage of using ℓ_1-norms inside the feature space, is that the method will adapt the complexity to the problem, by only selecting the right order of complexity from exponentially many features.

3.3 Designing New Decomposed Kernels

As shown in Section 5, the problem is well-behaved numerically and statistically if there is not too much correlation between the various feature maps Φ_v, $v \in V$. Thus, kernels such as the the all-subset Gaussian kernels may not be appropriate as each feature space contains the feature spaces of its ancestors[2]. Note that a strategy we could follow would be to remove some contributions of all ancestors by appropriate orthogonal projections. We now design specific kernels for which the feature space of each node is orthogonal to the feature spaces of its ancestors (for well-defined dot products).

[2]More precisely, this is true for the closures of these spaces of functions.

Spline kernels. In Eq. (10), we may chose, with $q = 2$:

$$\begin{aligned} k_{i0}(x_i, x_i') &= 1 \\ k_{i1}(x_i, x_i') &= x_i x_i' \\ k_{i2}(x_i, x_i') &= \min\{|x_i|, |x_i'|\}^2 (3\max\{|x_i|, |x_i'|\} - \min\{|x_i|, |x_i'|\})/6, \text{ if } x_i x_i' \geqslant 0 \\ &= 0, \text{ otherwise}, \end{aligned}$$

leading to tensor products of one-dimensional cubic spline kernels (Wahba, 1990; Gu, 2002). This kernel has the advantage of (a) being parameter free and (b) explicitly starting with linear features and essentially provides a convexification of multivariate additive regression splines (Friedman, 1991). Note that it may be more efficient here to use natural splines in the estimation method (Wahba, 1990) than using kernel matrices.

Hermite kernels. We can start from the following identity, valid for $\alpha < 1$ and from which the decomposition of the Gaussian kernel in Eq. (11) may be obtained (Szegö, 1981):

$$\sum_{j=0}^{\infty} \frac{\alpha^j}{j! 2^j} H_j(x_i) H_j(x_i') = (1-\alpha^2)^{-1/2} \exp\left(\frac{-2\alpha(x_i - x_i')^2}{1-\alpha^2} + \frac{(x_i^2 + (x_i')^2)\alpha}{1+\alpha} \right).$$

We can then define a sequence of kernel which also starts with linear kernels:

$$\begin{aligned} k_{i0}(x_i, x_i') &= H_0(x) H_0(x') = 1 \\ k_{ij}(x_i, x_i') &= \frac{\alpha^j}{2^j j!} H_j(x) H_j(x') \text{ for } j \in \{1, \ldots, q-1\} \\ k_{iq}(x_i, x_i') &= \sum_{j=q}^{\infty} \frac{\alpha^j}{j! 2^j} H_j(x_i) H_j(x_i'). \end{aligned}$$

Most kernels that we consider in this section (except the polynomial kernels) are universal kernels (Micchelli et al., 2006; Steinwart, 2002), that is, on a compact set of $\mathbb{R}^p$, their reproducing kernel Hilbert space is dense in $L^2(\mathbb{R}^p)$. This is the basis for the universal consistency results in Section 5.3. Moreover, some kernels such as the spline and Hermite kernels explicitly include the linear kernels inside their decomposition: in this situation, the sparse decomposition will start with linear features. In Section 5.3, we briefly study the universality of the kernel decompositions that we consider.

3.4 Kernels or Features?

In this paper, we emphasize the *kernel view*, i.e., we assume we are given a positive definite kernel (and thus a feature space) and we explore it using ℓ_1-norms. Alternatively, we could use the *feature view*, i.e., we would assume that we have a large structured set of features that we try to select from; however, the techniques developed in this paper assume that (a) each feature might be infinite-dimensional and (b) that we can sum all the local kernels efficiently (see in particular Section 4.2). Following the kernel view thus seems slightly more natural, but by no means necessary—see Jenatton et al. (2009) for a more general "feature view" of the problem.

In order to apply our optimization techniques in the feature view, as shown in Section 4, we simply need a specific *upper bound* on the kernel to be able to be computed efficiently. More precisely, we need to be able to compute $\sum_{w \in \mathrm{D}(t)} \left(\sum_{v \in \mathrm{A}(w) \cap \mathrm{D}(t)} d_v \right)^{-2} K_w$ for all $t \in V$, or an upper bound thereof, for appropriate weights (see Section 4.2 for further details).

3.5 Graph-Based Structured Regularization

Given $f \in \prod_{v \in V} \mathcal{F}_v$, the natural Hilbertian norm $\|f\|$ is defined through $\|f\|^2 = \sum_{v \in V} \|f_v\|^2$. Penalizing with this norm is efficient because summing all kernels k_v is assumed feasible in polynomial time and we can bring to bear the usual kernel machinery; however, it does not lead to sparse solutions, where many f_v will be exactly equal to zero, which we try to achieve in this paper.

We use the DAG to limit the set of active patterns to certain configurations, i.e., sets which are equal to their hulls, or equivalenty sets which contain all ancestors of their elements. If we were using a regularizer such as $\sum_{v \in V} \|f_v\|$ we would get sparse solutions, but the set of active kernels would be scattered throughout the graph and would not lead to optimization algorithms which are sub-linear in the number of vertices $|V|$.

All sets which are equal to their hull can be obtained by removing all the descendants of certain vertices. Indeed, the hull of a set I is characterized by the set of v, such that $\mathrm{D}(v) \subset I^c$, i.e., such that all descendants of v are in the complement I^c of I:

$$\mathrm{hull}(I) = \{v \in V, \ \mathrm{D}(v) \subset I^c\}^c.$$

Thus, if we try to estimate a set I such that $\mathrm{hull}(I) = I$, we thus need to determine which $v \in V$ are such that $\mathrm{D}(v) \subset I^c$. In our context, we are hence looking at selecting vertices $v \in V$ for which $f_{\mathrm{D}(v)} = (f_w)_{w \in \mathrm{D}(v)} = 0$. We thus consider the following structured block ℓ_1-norm defined on $\mathcal{F} = \mathcal{F}_1 \times \cdots \times \mathcal{F}_p$ as

$$\Omega(f) = \sum_{v \in V} d_v \|f_{\mathrm{D}(v)}\| = \sum_{v \in V} d_v \Big(\sum_{w \in \mathrm{D}(v)} \|f_w\|^2 \Big)^{1/2}, \tag{12}$$

where $(d_v)_{v \in V}$ are strictly positive weights. We assume that for all vertices but the sources of the DAG, we have $d_v = \beta^{\mathrm{depth}(v)}$ with $\beta > 1$, where $\mathrm{depth}(v)$ is the depth of node v, i.e., the length of the smallest path to the sources. We denote by $d_r \in (0, 1]$ the common weights to all sources. Other weights could be considered, in particular, weights inside the blocks $\mathrm{D}(v)$ (see, e.g. Jenatton et al., 2009), or weights that lead to penalties closer to the Lasso (i.e., $\beta < 1$), for which the effect of the DAG would be weaker. Note that when the DAG has no edges, we get back the usual block ℓ_1-norm with uniform weights d_r, and thus, the results presented in this paper (in particular the algorithm presented in Section 4.4 and non-asymptotic analysis presented in Section 5.2) can be applied to multiple kernel learning.

Penalizing by such a norm will indeed impose that some of the vectors $f_{\mathrm{D}(v)} \in \prod_{w \in \mathrm{D}(v)} \mathcal{F}_w$ are exactly zero, and we show in Section 5.1 that these are the only patterns we might get. We thus consider the following minimization problem[3]:

$$\min_{f \in \prod_{v \in V} \mathcal{F}_v, \, b \in \mathbb{R}} \ \frac{1}{n} \sum_{i=1}^{n} \ell\Big(y_i, \sum_{v \in V} \langle f_v, \Phi_v(x_i) \rangle + b \Big) + \frac{\lambda}{2} \Big(\sum_{v \in V} d_v \|f_{\mathrm{D}(v)}\| \Big)^2. \tag{13}$$

[3]Following Bach et al. (2004a) and Section 2, we consider the square of the norm, which does not change the regularization properties, but allow simple links with multiple kernel learning.

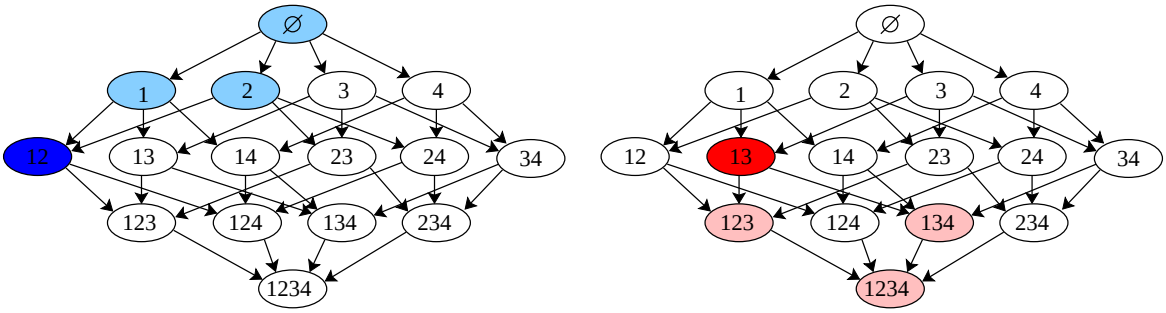

Figure 5: Directed acyclic graph of subsets of size 4: (left) a vertex (dark blue) with its ancestors (light blue), (right) a vertex (dark red) with its descendants (light red). By zeroing out weight vectors associated with descendants of several nodes, we always obtained a set of non-zero weights which contains all of its own ancestors (i.e., the set of non-zero weights is equal to its hull).

Our norm is a Hilbert space instantiation of the hierarchical norms recently introduced by Zhao et al. (2009). If all Hilbert spaces are finite dimensional, our particular choice of norms corresponds to an "ℓ_1-norm of ℓ_2-norms". While with uni-dimensional groups or kernels, the "ℓ_1-norm of ℓ_∞-norms" allows an efficient path algorithm for the square loss and when the DAG is a tree (Zhao et al., 2009), this is not possible anymore with groups of size larger than one, or when the DAG is not a tree (see Szafranski et al., 2008, for examples on two-layer hierarchies). In Section 4, we propose a novel algorithm to solve the associated optimization problem in polynomial time in the number of selected groups or kernels, for all group sizes, DAGs and losses. Moreover, in Section 5, we show under which conditions a solution to the problem in Eq. (13) consistently estimates the hull of the sparsity pattern.

4 Optimization

In this section, we give optimality conditions for the problems in Eq. (13), as well as optimization algorithms with polynomial time complexity in the number of selected kernels. In simulations, we consider total numbers of kernels up to 4^{256}, and thus such efficient algorithms that can take advantage of the sparsity of solutions are essential to the success of hierarchical multiple kernel learning (HKL).

4.1 Reformulation in terms of Multiple Kernel Learning

Following Rakotomamonjy et al. (2008), we can simply derive an equivalent formulation of Eq. (13). Using Cauchy-Schwarz inequality, we have that for all $\eta \in \mathbb{R}_+^V$ such that $\sum_{v \in V} d_v^2 \eta_v \leqslant 1$, a variational formulation of $\Omega(f)^2$ defined in Eq. (12):

$$\begin{aligned}\Omega(f)^2 &= \Big(\sum_{v \in V} d_v \|f_{\mathrm{D}(v)}\| \Big)^2 = \Big(\sum_{v \in V} (d_v \eta_v^{1/2}) \frac{\|f_{\mathrm{D}(v)}\|}{\eta_v^{1/2}} \Big)^2 \\ &\leqslant \sum_{v \in V} d_v^2 \eta_v \times \sum_{v \in V} \frac{\|f_{\mathrm{D}(v)}\|^2}{\eta_v} \leqslant \sum_{w \in V} \Big(\sum_{v \in \mathrm{A}(w)} \eta_v^{-1} \Big) \|f_w\|^2,\end{aligned}$$

with equality if and only if for all $v \in V$ $\eta_v = d_v^{-1}\|f_{\mathrm{D}(v)}\| \left(\sum_{w\in V} d_w \|f_{\mathrm{D}(w)}\|\right)^{-1} = \frac{d_v^{-1}\|f_{\mathrm{D}(v)}\|}{\Omega(f)}$.

We associate to the vector $\eta \in \mathbb{R}_+^V$, the vector $\zeta \in \mathbb{R}_+^V$ such that

$$\forall w \in V,\ \zeta_w(\eta)^{-1} = \sum_{v \in \mathrm{A}(w)} \eta_v^{-1}. \tag{14}$$

We use the natural convention that if η_v is equal to zero, then $\zeta_w(\eta)$ is equal to zero for all descendants w of v. We let denote $H = \{\eta \in \mathbb{R}_+^V,\ \sum_{v\in V} d_v^2 \eta_v \leqslant 1\}$ the set of allowed η and $Z = \{\zeta(\eta),\ \eta \in H\}$ the set of all associated $\zeta(\eta)$ for $\eta \in H$. The set H and Z are in bijection, and we can interchangeably use $\eta \in H$ or the corresponding $\zeta(\eta) \in Z$. Note that Z is in general not convex (unless the DAG is a tree, see Proposition 9 in Appendix A.1), and if $\zeta \in Z$, then $\zeta_w \leqslant \zeta_v$ for all $w \in \mathrm{D}(v)$, i.e., weights of descendant kernels are always smaller, which is consistent with the known fact that *kernels should always be selected after all their ancestors* (see Section 5.1 for a precise statement).

The problem in Eq. (13) is thus equivalent to

$$\min_{\eta\in H} \min_{f\in\prod_{v\in V}\mathcal{F}_v,\ b\in\mathbb{R}} \frac{1}{n}\sum_{i=1}^n \ell\Big(y_i, \sum_{v\in V}\langle f_v, \Phi_v(x_i)\rangle + b\Big) + \frac{\lambda}{2}\sum_{w\in V}\zeta_w(\eta)^{-1}\|f_w\|^2. \tag{15}$$

From Section 2, we know that at the optimum, $f_w = \zeta_w(\eta)\sum_{i=1}^n \alpha_i \Phi_w(x_i) \in \mathcal{F}_w$, where $\alpha \in \mathbb{R}^n$ are the dual parameters associated with the single kernel learning problem in Proposition 1, with kernel matrix $\sum_{w\in V}\zeta_w(\eta)K_w$.

Thus, the solution is entirely determined by $\alpha \in \mathbb{R}^n$ and $\eta \in H \subset \mathbb{R}^V$ (and its corresponding $\zeta(\eta) \in Z$). We also associate to α and η the corresponding functions f_w, $w \in V$, and optimal constant b, for which we can check optimality conditions. More precisely, we have (see proof in Appendix A.4):

Proposition 2 (Dual problem for HKL) *The convex optimization problem in Eq. (13) has the following dual problem:*

$$\max_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0} -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2}\max_{\eta\in H}\sum_{w\in V}\zeta_w(\eta)\alpha^\top\widetilde{K}_w\alpha. \tag{16}$$

Moreover, at optimality, $\forall w \in V, f_w = \zeta_w(\eta)\sum_{i=1}^n \alpha_i\Phi_w(x_i)$ and $b = b^*\left(\sum_{w\in V}\zeta_w(\eta)K_w\alpha\right)$, with η attaining, given α, the maximum of $\sum_{w\in V}\zeta_w(\eta)\alpha^\top\widetilde{K}_w\alpha$.

Proposition 3 (Optimality conditions for HKL) *Let $(\alpha,\eta) \in \mathbb{R}^n \times H$, such that $1_n^\top\alpha = 0$. Define functions $f \in \mathcal{F}$ through $\forall w \in V, f_w = \zeta_w(\eta)\sum_{i=1}^n \alpha_i\Phi_w(x_i)$ and $b = b^*\left(\sum_{w\in V}\zeta_w(\eta)K_w\alpha\right)$ the corresponding constant term. The vector of functions f is optimal for Eq. (13), if and only if :*

(a) given $\eta \in H$, the vector α is optimal for the single kernel learning problem with kernel matrix $K = \sum_{w\in V}\zeta_w(\eta)K_w$,

(b) given α, $\eta \in H$ maximizes

$$\sum_{w\in V}\Big(\sum_{v\in\mathrm{A}(w)}\eta_v^{-1}\Big)^{-1}\alpha^\top\widetilde{K}_w\alpha = \sum_{w\in V}\zeta_w(\eta)\alpha^\top\widetilde{K}_w\alpha. \tag{17}$$

Moreover, as shown in Appendix A.4, the total duality gap can be upperbounded as the sum of the two separate duality gaps for the two optimization problems, which will be useful in Section 4.2 for deriving sufficient conditions of optimality (see Appendix A.4 for more details):

$$\mathrm{gap}_{\mathrm{kernel}}\Big(\sum_{w\in V}\zeta_w(\eta)\widetilde{K}_w,\alpha\Big)+\frac{\lambda}{2}\mathrm{gap}_{\mathrm{weights}}\big((\alpha^\top\widetilde{K}_w\alpha)_{w\in V},\eta\big), \tag{18}$$

where $\mathrm{gap}_{\mathrm{weights}}$ corresponds to the duality gap of Eq. (17). Note that in the case of "flat" regular multiple kernel learning, where the DAG has no edges, we obtain back usual optimality conditions (Rakotomamonjy et al., 2008; Pontil and Micchelli, 2005).

Following a common practice for convex sparse problems (Lee et al., 2007; Roth and Fischer, 2008), we will try to solve a small problem where we assume we know the set of v such that $\|f_{\mathrm{D}(v)}\|$ is equal to zero (Section 4.3). We then need (a) to check that variables in that set may indeed be left out of the solution, and (b) to propose variables to be added if the current set is not optimal. In the next section, we show that this can be done in polynomial time although the number of kernels to consider leaving out is exponential (Section 4.2).

Note that an alternative approach would be to consider the regular multiple kernel learning problem with additional linear constraints $\zeta_{\pi(v)}\geqslant\zeta_v$ for all non-sources $v\in V$. However, it would not lead to the analysis through sparsity-inducing norms outlined in Section 5 and might not lead to polynomial-time algorithms.

4.2 Conditions for Global Optimality of Reduced Problem

We consider a subset W of V which is equal to its hull—as shown in Section 5.1, those are the only possible active sets. We consider the optimal solution f of the reduced problem (on W), namely,

$$\min_{f_W\in\prod_{v\in W}\mathcal{F}_v,\,b\in\mathbb{R}}\ \frac{1}{n}\sum_{i=1}^n\ell\Big(y_i,\sum_{v\in W}\langle f_v,\Phi_v(x_i)\rangle+b\Big)+\frac{\lambda}{2}\Big(\sum_{v\in W}d_v\|f_{\mathrm{D}(v)\cap W}\|\Big)^2, \tag{19}$$

with optimal primal variables f_W, dual variables $\alpha\in\mathbb{R}^n$ and optimal pair (η_W,ζ_W). From these, we can construct a full solution f to the problem, as $f_{W^c}=0$, with $\eta_{W^c}=0$. That is, we keep α unchanged and add zeros to η_W.

We now consider necessary conditions and sufficient conditions for this augmented solution to be optimal with respect to the full problem in Eq. (13). We denote by $\Omega(f)=\sum_{v\in W}d_v\|f_{D(v)\cap W}\|$ the optimal value of the norm for the reduced problem.

Proposition 4 (Necessary optimality condition) *If the reduced solution is optimal for the full problem in Eq. (13) and all kernels indexed by W are active, then we have:*

$$\max_{t\in\mathrm{sources}(W^c)}\frac{\alpha^\top\widetilde{K}_t\alpha}{d_t^2}\leqslant\Omega(f)^2. \tag{20}$$

Proposition 5 (Sufficient optimality condition) *If*

$$\max_{t\in\mathrm{sources}(W^c)}\sum_{w\in\mathrm{D}(t)}\frac{\alpha^\top\widetilde{K}_w\alpha}{(\sum_{v\in\mathrm{A}(w)\cap\mathrm{D}(t)}d_v)^2}\leqslant\Omega(f)^2+2\varepsilon/\lambda, \tag{21}$$

then the total duality gap in Eq. (18) is less than ε.

The proof is fairly technical and can be found in Appendix A.5; this result constitutes the main technical result of the paper: it essentially allows to design an algorithm for solving a large optimization problem over exponentially many dimensions in polynomial time. Note that when the DAG has no edges, we get back regular conditions for unstructured MKL—for which Eq. (20) is equivalent to Eq. (21) for $\varepsilon = 0$.

The necessary condition in Eq. (20) does not cause any computational problems as the number of sources of W^c, i.e., the cardinal of $\mathrm{sources}(W^c)$, is upper-bounded by $|W|$ times the maximum out-degree of the DAG.

However, the sufficient condition in Eq. (21) requires to sum over all descendants of the active kernels, which is impossible without special structure (namely exactly being able to compute that sum or an upperbound thereof). Here, we need to bring to bear the specific structure of the full kernel k. In the context of directed grids we consider in this paper, if d_v can also be decomposed as a product, then $\sum_{v \in \mathrm{A}(w) \cap \mathrm{D}(t)} d_v$ can also be factorized, and we can compute the sum over all $v \in \mathrm{D}(t)$ in linear time in p. Moreover, we can cache the sums

$$\breve{K}_t = \sum_{w \in \mathrm{D}(t)} \left(\sum_{v \in \mathrm{A}(w) \cap \mathrm{D}(t)} d_v \right)^{-2} \widetilde{K}_w$$

in order to save running time in the active set algorithm presented in Section 4.4. Finally, in the context of directed grids, many of these kernels are either constant across iterations, or change slightly; that is, they are product of sums, where most of the sums are constant across iterations, and thus computing a new cached kernel can be considered of complexity $O(n^2)$, independent of the DAG and of W.

4.3 Dual Optimization for Reduced or Small Problems

In this section, we consider solving Eq. (13) for DAGs V (or active set W) of small cardinality, i.e., for (very) small problems or for the reduced problems obtained from the algorithm presented in Figure 6 from Section 4.4.

When kernels k_v, $v \in V$, have low-dimensional feature spaces, either by design (e.g., rank one if each node of the graph corresponds to a single feature), or after a low-rank decomposition such as a singular value decomposition or an incomplete Cholesky factorization (Fine and Scheinberg, 2001; Bach and Jordan, 2005), we may use a "primal representation" and solve the problem in Eq. (13) using generic optimization toolboxes adapted to conic constraints (see, e.g., Grant and Boyd, 2008). With high-dimensional feature spaces, in order to reuse existing optimized supervised learning code and use high-dimensional kernels, it is preferable to use a "dual optimization". Namely, we follow Rakotomamonjy et al. (2008), and consider for $\zeta \in Z$, the function

$$B(\zeta) = G(K(\zeta)) = \min_{f \in \prod_{v \in V} \mathcal{F}_v,\ b \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(y_i, \sum_{v \in V} \langle f_v, \Phi_v(x_i) \rangle + b \right) + \frac{\lambda}{2} \sum_{w \in V} \zeta_w^{-1} \|f_w\|^2,$$

which is the optimal value of the single kernel learning problem with kernel matrix $\sum_{w \in V} \zeta_w K_w$. Solving Eq. (15) is equivalent to minimizing $B(\zeta(\eta))$ with respect to $\eta \in H$.

If the Fenchel conjugate of the loss is strictly convex (i.e., square loss, logistic loss, Hüber loss, 2-norm support vector regression), then the function B is differentiable—because the dual problem in Eq. (3) has a unique solution α (Bonnans and Shapiro, 2000). When the Fenchel conjugate is

not strictly convex, a ridge (i.e., positive diagonal matrix) may be added to the kernel matrices, which has the exact effect of smoothing the loss—see, e.g., Lemaréchal and Sagastizábal (1997) for more details on relationships between smoothing and adding strongly convex functions to the dual objective function.

Moreover, the function $\eta \mapsto \zeta(\eta)$ is differentiable on $(\mathbb{R}_+^*)^V$, but not at any points η such that one η_v is equal to zero. Thus, the function $\eta \mapsto B[\zeta((1-\varepsilon)\eta + \frac{\varepsilon}{|V|}d^{-2})]$, where d^{-2} is the vector with elements d_v^{-2}, is differentiable if $\varepsilon > 0$, and its derivatives can simply be obtained from the chain rule. In simulations, we use $\varepsilon = 10^{-3}$; note that adding this term is equivalent to smoothing the norm $\Omega(f)$ (i.e., make it differentiable), while retaining its sparsity-inducing properties (i.e., some of the optimal η will still be exactly zero).

We can then use the same projected gradient descent strategy as Rakotomamonjy et al. (2008) to minimize it. The overall complexity of the algorithm is then proportional to $O(|V|n^2)$—to form the kernel matrices—added to the complexity of solving a single kernel learning problem—typically between $O(n^2)$ and $O(n^3)$, using proper kernel classification/regression algorithms (Vishwanathan et al., 2003; Loosli et al., 2005). Note that we could follow the approach of Chapelle and Rakotomamonjy (2008) and consider second-order methods for optimizing with respect to η.

4.4 Kernel Search Algorithm

We now present the detailed algorithm which extends the search algorithm of Lee et al. (2007) and Roth and Fischer (2008). Note that the kernel matrices are never all needed explicitly, i.e., we only need them (a) explicitly to solve the small problems (but we need only a few of those) and (b) implicitly to compute the necessary condition in Eq. (20) and the sufficient condition in Eq. (21), which requires to sum over all kernels which are not selected, as shown in Section 4.2.

The algorithm works in two phases: first the (local) necessary condition is used to check optimality of the solution and add variables; when those are added, the augmented reduced problem must include the new variable into the active set. Once the necessary condition is fulfilled, we use the sufficient condition, which essentially sums over all non selected kernels and makes sure that if some information is present further away in the graph, it will indeed be selected. See Figure 6 for details[4].

The algorithm presented in Figure 6 will stop either when the duality gap is less than 2ε or when the maximal number of kernels Q has been reached. That is, our algorithm does not always yield a solution which is provably approximately optimal. In practice, when the weights d_v increase with the depth of v in the DAG (which we use in simulations), the provably small duality gap generally occurs before we reach a problem larger than Q (however, we cannot make sharp statements). Note that some of the iterations only increase the size of the active sets to check the sufficient condition for optimality. Forgetting those would not change the solution as we add kernels with zero weights; however, in this case, we would not be able to actually certify that we have an 2ε-optimal solution (see Figure 7 for an example of these two situations). Note that because of potential overfitting issues, settings of the regularization parameter λ with solutions having more than n active kernels are likely to have low predictive performance. Therefore, we may expect the algorithm to be useful in practice with moderate values of Q.

[4]Matlab/C code for least-squares regression and logistic regression may be downloaded from the author's website.

Input: Kernel matrices $K_v \in \mathbb{R}^{n \times n}$, weights d_v, $v \in V$, maximal gap ε, maximal number of kernels Q.
Algorithm:

1. Initialization: active set $W = \varnothing$, cache kernel matrices $\breve{K}_w$, $w \in \mathrm{sources}(W^c)$
2. Compute (α, η) solutions of Eq. (19), obtained using Section 4.3 (with gap ε)
3. While necessary condition in Eq. (20) is not satisfied and $|W| \leqslant Q$
   a. Add violating kernel in $\mathrm{sources}(W^c)$ to W
   b. Compute (α, η) solutions of Eq. (19), obtained using Section 4.3 (with gap ε)
   c. Update cached kernel matrices $\breve{K}_w$, $w \in \mathrm{sources}(W^c)$
4. While sufficient condition in Eq. (21) is not satisfied and $|W| \leqslant Q$
   a. Add violating kernel in $\mathrm{sources}(W^c)$ to W
   b. Compute (α, η) solutions of Eq. (19), obtained using Section 4.3 (with gap ε)
   c. Update cached kernel matrices $\breve{K}_w$, $w \in \mathrm{sources}(W^c)$

Output: W, α, η, constant term b

Figure 6: Kernel search algorithm for hierarchical kernel learning. The algorithm stops either when the duality gap is provably less than 2ε, either when the maximum number of active kernels has been achieved; in the latter case, the algorithm may or may not have reached a 2ε-optimal solution (i.e., a solution with duality gap less than 2ε).

Running-time complexity. Let D be the maximum out-degree (number of children) in the graph, κ be the complexity of evaluating the sum in the sufficient condition in Eq. (21) (which usually takes constant time), and $R = |W|$ the number of selected kernels (the number is the size of the active set W). Assuming $O(n^3)$ for the single kernel learning problem, which is conservative (see, e.g. Vishwanathan et al., 2003; Loosli et al., 2005, for some approaches), solving all reduced problems has complexity $O(Rn^3)$. Computing all cached matrices has complexity $O(\kappa n^2 \times RD)$ and computing all necessary/sufficient conditions has complexity $O(n^2 \times R^2 D)$. Thus, the total complexity is $O(Rn^3 + \kappa n^2 RD + n^2 R^2 D)$. Thus, in the case of the directed p-grid, we get $O(Rn^3 + n^2 R^2 p)$. Note that the kernel search algorithm is also an efficient algorithm for unstructured MKL, for which we have complexity $O(Rn^3 + n^2 R^2 p)$. Note that gains could be made in terms of scaling with respect to n by using better kernel machine codes with complexity between $O(n^2)$ and $O(n^3)$ (Vishwanathan et al., 2003; Loosli et al., 2005). Note that while the algorithm has polynomial complexity, some work is still needed to make it scalable for more than a few hundreds variables, in particular because of the memory requirements of $O(Rpn^2)$. In order to save storing requirements for the cached kernel matrices, low-rank decompositions might be useful (Fine and Scheinberg, 2001; Bach and Jordan, 2005).

5 Theoretical Analysis in High-Dimensional Settings

In this section, we consider the consistency of kernel selection for the norm $\Omega(f)$ defined in Section 3. In particular, we show formally in Section 5.1 that the active set is always equal to its hull, and provide in Section 5.2 conditions under which the hull is consistently estimated in low and high-dimensional settings, where the cardinality of V may be large compared to the number of observa-

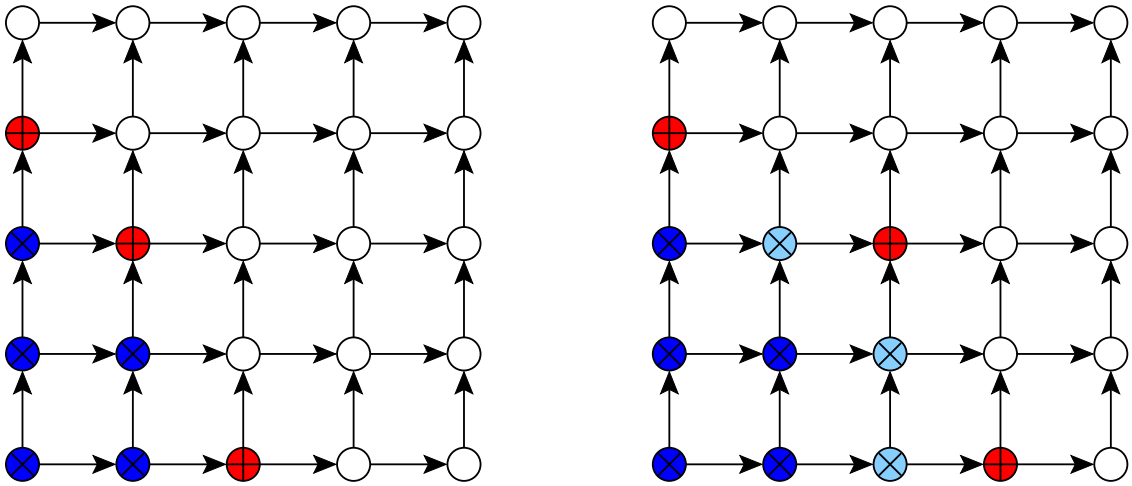

Figure 7: Example of active sets for the kernel search algorithms: (left) first phase, when checking necessary conditions, the dark blue nodes ($\times$) are the active kernels (non-zero η), and the red $+$ are the sources of the complement, which may be added at the next iteration; (right) second phase, when checking sufficient conditions, the dark blue nodes ($\times$) are the active kernels (non-zero η), the light blue nodes ($\times$) are the kernels with zero weights but are here just to check optimality conditions, and the red nodes ($+$) are the sources of the complement, which may be added at the next iteration.

tions. Throughout this section, we denote by $\hat{f}$ any minimizer of Eq. (13) and $\hat{W} = \{v \in V, \hat{f}_v \neq 0\}$ the set of selected kernels.

5.1 Allowed Patterns

We now show that under certain assumptions any solution of Eq. (13) will have a nonzero pattern which is equal to its hull, i.e., the set $\hat{W} = \{v \in V, \hat{f}_v \neq 0\}$ must be such that $\hat{W} = \bigcup_{w \in \hat{W}} \mathrm{A}(w)$— see Jenatton et al. (2009) for a more general result with overlapping groups without the DAG structure and potentially low-rank kernels:

Theorem 6 (Allowed patterns) *Assume that all kernel matrices are invertible. Then the set of zeros $\hat{W}$ of any solution $\hat{f}$ of Eq. (13) is equal to its hull.*

Proof Since the dual problem in Eq. (16) has a strictly convex objective function on the hyperplane $\alpha^\top 1_n = 0$, the minimum in $\alpha \in \mathbb{R}^n$ is unique. Moreover, we must have $\alpha \neq 0$ as soon as the loss functions φ_i are not all identical. Since $\|f_w\|^2 = \zeta_w^2 \alpha^\top \widetilde{K}_w \alpha$ for some $\zeta \in Z$, and all $\alpha^\top \widetilde{K}_w \alpha > 0$ (by invertibility of K_w and $\alpha^\top 1_n = 0$), we get the desired result, from the sparsity pattern of the vector $\zeta \in \mathbb{R}^V$, which is always equal to its hull. ■

As shown above, the sparsity pattern of the solution of Eq. (13) will be equal to its hull, and thus we can only hope to obtain consistency of the hull of the pattern, which we consider in the next sections. In Section 5.2, we provide a sufficient condition for optimality, whose weak form tends to be also necessary for consistent estimation of the hull; these results extend the one for the Lasso and the group Lasso (Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007; Wainwright, 2009; Bach, 2008a).

5.2 Hull Consistency Condition

For simplicity, we consider the square loss for regression and leave out other losses presented in Section 2.1 for future work. Following Bach (2008a), we consider a random design setting where the pairs $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$ are sampled from *independent and identical distributions*. We make the following assumptions on the DAG, the weights of the norm and the underlying joint distribution of $(\Phi_v(X))_{v \in V}$ and Y. These assumptions rely on *covariance operators*, which are the tools of choice for analyzing supervised and unsupervised learning techniques with reproducing kernel Hilbert spaces (see Bach, 2008a; Fukumizu et al., 2007; Harchaoui et al., 2008, for a introduction to the main concepts which are used in this paper). We let denote $\mathbf{\Sigma}$ the joint covariance operator for the kernel $k(x, y)$ defined by blocks corresponding to the decomposition indexed by V. We make the following assumptions:

(A0) *Weights of the DAG*: Each of the $\mathrm{num}(V)$ strongly connected components of V has a unique source; the weights of the sources are equal to $d_r \in (0, 1]$, while all other weights are equal to $d_v = \beta^{\mathrm{depth}(v)}$ with $\beta > 1$. The maximum out-degree (number of children) of the DAG is less than $\deg(V) - 1$.

(A1) *Sparse non-linear model*: $\mathbb{E}(Y|X) = \sum_{w \in \mathbf{W}} \langle \mathbf{f}_w(X) + \mathbf{b}$ with $\mathbf{W} \subset V$, $\mathbf{f}_w \in \mathcal{F}_w$, $w \in \mathbf{W}$, and $\mathbf{b} \in \mathbb{R}$; the conditional distribution of $Y|X$ is Gaussian with variance $\boldsymbol{\sigma}^2 > 0$. The set $\mathbf{W}$ is equal to its hull, and for each $w \in \mathbf{W}$, $\mathbf{f}_{\mathrm{D}(w) \cap \mathbf{W}} \neq 0$ (i.e., the hull of the non zero functions is actually $\mathbf{W}$).

(A2) *Uniformly bounded inputs*: for all $v \in V$, $\|\Phi_v(X)\| \leqslant 1$ almost surely, i.e., $k_v(X, X) \leqslant 1$.

(A3) *Compacity and invertibility of the correlation operator on the relevant variables*: The joint correlation operator $\boldsymbol{C}$ of $(\Phi(x_v))_{v \in V}$ (defined with appropriate blocks $\boldsymbol{C}_{vw}$) is such that $\boldsymbol{C}_{\mathbf{WW}}$ is compact and invertible (with smallest eigenvalue $\kappa = \lambda_{\min}(\boldsymbol{C}_{\mathbf{WW}}) > 0$).

(A4) *Smoothness of predictors*: For each $w \in \mathbf{W}$, there exists $\mathbf{h}_w \in \mathcal{F}_w$ such that $\mathbf{f}_w = \mathbf{\Sigma}_{ww} \mathbf{h}_w$ and $\|\mathbf{h}_w\| \leqslant 1$.

(A5) *Root-summability of eigenvalues of covariance operators*: For each $w \in \mathbf{W}$, the sum of the square roots of the eigenvalues of $\mathbf{\Sigma}_{ww}$ is less than a constant $C_{1/2}$.

When the Hilbert spaces all have finite dimensions, covariance operators reduce to covariance matrices, and Assumption **(A3)** reduces to the invertibility of the correlation matrix $\boldsymbol{C}_{\mathbf{WW}}$ (as it is always compact) and thus of the covariance matrix $\mathbf{\Sigma}_{\mathbf{WW}}$, while **(A4)** and **(A5)** are always satisfied. These assumptions are discussed by Bach (2008a) in the context of multiple kernel learning, which is essentially our framework with a trivial DAG with no edges (and as many connected components as kernels). Note however that Assumption **(A4)** is slightly stronger than the one used by Bach (2008a) and that we derive here non asymptotic results, while Bach (2008a) was considering only asymptotic results.

For K a subset of V, we denote by $\Omega_K(f_K) = \sum_{v \in K} d_v \|f_{\mathrm{D}(v) \cap K}\|$, the norm reduced to the functions in K and by Ω_K^* its *dual norm* (Boyd and Vandenberghe, 2003; Rockafellar, 1970), defined as $\Omega_K^*(g_K) = \max_{\Omega_K(f_K) \leqslant 1} \langle g_K, f_K \rangle$. We consider $\mathbf{s}_{\mathbf{W}} \in (\mathcal{F}_v)_{v \in \mathbf{W}}$, defined through

$$\forall w \in \mathbf{W},\ \mathbf{s}_w = \Big(\sum_{v \in \mathrm{A}(w)} d_v \|\mathbf{f}_{\mathrm{D}(v)}\|^{-1} \Big) \mathbf{h}_w.$$

When the DAG has no edges, i.e., for the regular group Lasso, we get back similar quantities than the ones obtained by Bach (2008a); if in addition, the feature spaces are all uni-dimensional, we get the vector of signs of the relevant variables, recovering the Lasso conditions (Zhao and Yu, 2006; Zou, 2006; Yuan and Lin, 2007; Wainwright, 2009). The following theorem shows that if the consistency condition in Eq. (22) is satisfied, then we can upperbound the probability of incorrect hull selection (see proof in Appendix B):

Theorem 7 (Sufficient condition for hull consistency) *Assume* **(A0-5)** *and*

$$\Omega^*_{\mathbf{W}^c}\Big[\mathrm{Diag}(\mathbf{\Sigma}^{1/2}_{ww})_{\mathbf{W}^c}\boldsymbol{C}_{\mathbf{W}^c\mathbf{W}}\boldsymbol{C}^{-1}_{\mathbf{W}\mathbf{W}}\mathbf{s}_{\mathbf{W}}\Big] \leqslant 1-\eta, \tag{22}$$

with $\eta > 0$*; let* $\nu = \min_{w\in\mathbf{W}} \|\mathrm{Diag}(\mathbf{\Sigma}_{vv})_{\mathrm{D}(w)}\mathbf{f}_{\mathrm{D}(w)}\|$ *and* $\omega = \Omega(\mathbf{f})d_r^{-2}$*. Let*

$$\gamma(V) = \frac{4\log(2\mathrm{num}(V))}{(1-\beta^{-1})^2} + \frac{4\log\deg(V)}{(\log\beta)^3}.$$

Choose $\mu = \lambda\Omega(\mathbf{f})d_r \in \Big[\frac{2\boldsymbol{\sigma}\gamma(V)^{1/2}}{n^{1/2}}, \frac{c_1}{\omega^{11/2}|\mathbf{W}|^{7/2}}\Big]$*. The probability of incorrect hull selection is upper-bounded by:*

$$\exp\Big(-\frac{\mu^2 n}{8\boldsymbol{\sigma}^2}\Big) + \exp\Big(-c_2\frac{\mu n}{\omega^3|\mathbf{W}|^3}\Big) + \exp\Big(-c_3\frac{\mu^{3/2}n}{\boldsymbol{\sigma}^2\omega^7|\mathbf{W}|^4}\Big), \tag{23}$$

where c_1*,* c_2*,* c_3 *are positive monomials in* κ*,* ν*,* η *and* $C^{-1}_{1/2}$*.*

The previous theorem is the main theoretical contribution of this paper. It is a non-asymptotic result which we comment on in the next paragraphs. The proof relies on novel concentration inequalities for empirical covariance operators and for structured norms, which may be useful in other settings (see results in Appendices B.2, B.3 and B.4). Note that the last theorem is not a consequence of similar results for flat multiple kernel learning or group Lasso (Bach, 2008a; Nardi and Rinaldo, 2008; Lounici et al., 2009), because the groups that we consider are overlapping. Moreover, the last theorem shows that we can indeed estimate the correct hull of the sparsity pattern if the sufficient condition is satisfied. In particular, if we can make the groups such that the between-group correlation is as small as possible, we can ensure correct hull selection.

Low-dimensional settings. When the DAG is assumed fixed (or in fact only the number of connected components $\mathrm{num}(V)$ and the maximum out-degree $\deg(V)$) and n tends to $+\infty$, the probability of incorrect hull selection tends to zero as soon as $\lambda n^{1/2}$ tends to $+\infty$ and λ tends to zero, and the convergence is exponentially fast in λn.

High-dimensional settings. When the DAG is large compared to n, then, the previous theorem leads to a consistent estimation of the hull, if the interval defining μ is not empty, i.e., $n \geqslant 4\boldsymbol{\sigma}^2\gamma(V)\omega^{11}|\mathbf{W}|^7 c_1^{-2}$. Since $\gamma(V) = O(\log(\mathrm{num}(V)) + \log(\deg(V)))$, this implies that we may have correct hull selection in situations where $n = O(\log(\mathrm{num}(V)) + \log(\deg(V)))$. We may thus have an exponential number of connected components and an exponential out-degree, with no constraints on the maximum depth of the DAG (it could thus be infinite).

Here, similar scalings could be obtained with a weighted ℓ_1-norm (with the same weights $\beta^{\mathrm{depth}(v)}$; however, such a weighted Lasso might select kernels which are far from the roor and would not be amenable to an efficient active set algorithm.

Multiple kernel learning (group Lasso). In this situation, we have a DAG with p connected components (one for each kernel), and zero out-degree (i.e., $\deg(V) = 1$), leading to $\gamma(V) = O((\log p)^{1/2})$, a classical non-asymptotic result in the unstructured settings for finite-dimensional groups (Nardi and Rinaldo, 2008; Wainwright, 2009; Lounici et al., 2009), but novel for the multiple kernel learning framework, where groups are infinite-dimensional Hilbert spaces. Note that the proof techniques would be much simpler and the result sharper in terms of power of $|\mathbf{W}|$ and ω with finite-dimensional groups and with the assumption of invertibility of $\mathbf{\Sigma_{WW}}$ and/or fixed design assumptions. Finally, Theorem 7 also applies for a modified version of the elastic net (Zou and Hastie, 2005), where the ℓ_2-norm is added to the sum of block ℓ_1 norm—by considering a single node with the null kernel connected to all other kernels.

Non linear variable selection. For the power set and the directed grids that we consider for non-linear variable selection in Section 3.2, we have $\mathrm{num}(V) = 1$ and $\deg(V) = p$ where p is the number of variables, and thus $\gamma(V) = O(\log p) = O(\log\log|V|)$, i.e., we may have exponentially many variables to choose non-linearly from, or a *doubly* exponential number of kernels to select from.

Trade-off for weight β. Intuitively, since the weight on the norm $\|f_{\mathrm{D}(v)}\|$ is equal to $\beta^{\mathrm{depth}(v)}$, the greater the β the stronger the prior towards selecting nodes close to the sources. However, if β is too large, the prior might be too strong to allow selecting nodes away from the sources.

This can be illustrated in the bound provided in Theorem 7. The constant $\gamma(V)$ is a decreasing function of β, and thus having a large β, i.e., a large penalty on the deep vertices, we decrease the lower bound of allowed regularization parameters μ and thus increase the probability of correct hull selection (far away vertices are more likely to be left out). However, since $\Omega(\mathbf{f})$ is a rapidly increasing function of β, the upper bound decreases, i.e., if we penalize too much, we would start losing some of the deeper relevant kernels. Finally, it is worth noting that if the constant β tend to infinity slowly with n, then we could always consistently estimate the depth of the hull, i.e., the optimal interaction complexity. Detailed results are the subject of ongoing work.

Results on estimation accuracy and predictive performance. In this paper, we have focused on the simpler results of hull selection consistency, which allow simple assumptions. It is however of clear interest of following the Lasso work on estimation accuracy and predictive performance (Bickel et al., 2009) and extend it to our structured setting. In particular, the rates of convergence should also depend on the cardinal of the active set $|\mathbf{W}|$ and not on the cardinality of the DAG $|V|$.

Enhancing consistency condition. The sufficient condition in Eq. (22) states that low correlation between relevant and irrelevant feature spaces leads to good model selection. As opposed to unstructured situations, such low correlation may be enhanced with proper hierarchical whitening of the data, i.e., for all $v \in V$, we may project $(\Phi_v(x_i))_{i=1,\dots,n}$ to the orthogonal of all ancestor vectors $(\Phi_w(x_i))_{i=1,\dots,n}$, $w \in \mathrm{A}(v)$. This does not change the representation power of our method but simply enhances its statistical consistency.

Moreover, Assumption (**A3**) is usually met for all the kernel decompositions presented in Section 3.2, except the all-subset Gaussian kernel (because each feature space of each node contains the

feature spaces associated with its parents). However, by the whitening procedure outlined above, similar results than Theorem 7 might be obtained. Besides, if the *original variables* used to define the kernel decompositions presented in Section 3.2 are independent, then the consistency condition in Eq. (22) is always met except for the all-subset Gaussian kernel; again, a pre-whitening procedure might solve the problem in this case.

Necessary consistency condition. We also have a necessary condition which is a weak form of the sufficient condition in Eq. (22)—the proof follows closely the one for the unstructured case from Bach (2008a):

Proposition 8 (Necessary condition for hull consistency) *Assume* **(A1-3)** *and V is fixed, with n tending to $+\infty$. If there is a sequence of regularization parameters λ such that both the prediction function and the hull of the active kernels is consistently estimated, then we have*

$$\Omega^*_{\mathbf{W}^c}\big[\operatorname{Diag}(\boldsymbol{\Sigma}^{1/2}_{ww})_{\mathbf{W}^c}\boldsymbol{C}_{\mathbf{W}^c\mathbf{W}}\boldsymbol{C}^{-1}_{\mathbf{W}\mathbf{W}}\mathbf{s}_{\mathbf{W}}\big] \leqslant 1. \tag{24}$$

The conditions in Eq. (22) and Eq. (24) make use of the dual norm, but we can loosen them using lower and upper bounds on these dual norms: some are computable in polynomial time, like the ones used for the active set algorithm presented in Section 4.4 and more detailed in Appendix B.7. However, we can obtain simpler bounds which require to look over the entire DAG; we obtain by lowerbounding $\|f_{\mathrm{D}(v)}\|$ by $\|f_v\|$ and upperbounding it by $\sum_{w\in\mathrm{D}(v)}\|f_w\|$ in the definition of $\Omega(f)$, for $g \in \mathcal{F}$:

$$\max_{w\in\mathbf{W}^c}\frac{\|g_w\|}{\sum_{v\in\mathrm{A}(w)\cap\mathbf{W}^c} d_v} \leqslant \Omega^*_{\mathbf{W}^c}(g_{\mathbf{W}^c}) \leqslant \max_{w\in\mathbf{W}^c}\frac{\|g_w\|}{d_w}.$$

The lower and upper bounds are equal when the DAG is trivial (no edges), and we get back the usual weighted ℓ_∞-ℓ_2 norm $\max_{w\in\mathbf{W}^c}\frac{\|g_w\|}{d_w}$.

5.3 Universal Consistency

In this section, we briefly discuss the universal consistency properties of our method when used for non-linear variable selection: do the kernel decompositions presented in Section 3.2 allow the estimation of arbitrary functions? The main rationale behind using all subsets of variables rather than only singletons is that most non-linear functions may not be expressed as a sum of functions which depend only on one variable—what regular MKL (Bach et al., 2004a) and SPAM (Ravikumar et al., 2008) would use. All subsets are thus required to allow universal consistency, i.e., to be able to approach any possible predictor function.

Our norm $\Omega(\mathbf{f})$ is equivalent to a weighted Hilbertian norm, i.e.:

$$\sum_{v\in V} d_v\|f_v\|^2 \leqslant \Omega(f)^2 \leqslant |V|\sum_{w\in V}\Big(\sum_{v\in\mathrm{A}(w)} d_v\Big)\|f_w\|^2.$$

Therefore, the usual RKHS balls associated to the universal kernels we present in Section 3.2 are contained in the ball of our norms, hence we obtain universal consistency (Steinwart, 2002; Micchelli et al., 2006) in low-dimensional settings when p is small. A more detailed and refined analysis that takes into account the sparsity of the decomposition and convergence rates is out of the scope of this paper, in particular for the different regimes for p, q and n.

6 Simulations

In this section, we report simulation experiments on synthetic datasets and datasets from the UCI repository. Our goals here are (a) to compare various kernel-based approaches to least-squares regression from the *same* kernel, (b) to compare the various kernel decompositions presented in Section 3.2 within our HKL framework, and (c) to compare predictive performance with non-kernel-based methods—more simulations may be found in earlier work (Bach, 2008b).

6.1 Compared Methods

In this section, we consider various nonparametric methods for non-linear predictions. Some are based on the kernel decompositions defined in Section 3.2. Non-kernel based methods were chosen among methods with some form of variable selection capabilities. All these methods were used with two loops of 10-fold cross-validation to select regularization parameters and hyperparameters (in particular β). All results are averaged over 10 replications (medians, upper and lower quartiles are reported).

Hierarchical kernel learning (HKL). We use the algorithm presented in Section 4.4 with the kernel decompositions presented in Section 3.2, i.e., Hermite polynomials ("Hermite"), spline kernels ("spline") and all-subset Gaussian kernels ("Gaussian").

Multiple kernel learning (MKL). We use the algorithm presented in Section 4.4 with the kernel decompositions presented in Section 3.2, but limited to kernels of depth one, which corresponds to sparse generalized additive models.

Constrained forward selection (greedy). Given a kernel decomposition with rank one kernels, we consider a forward selection approach that satisfies the same constraint that we impose in our convex framework.

Single kernel learning (L_2). When using the full decomposition (which is equivalent to summing all kernels or penalizing by an ℓ_2-norm) we can use regular single kernel learning.

Generalized Lasso (Glasso). Given the same kernel matrix as in the previous method, Roth (2004) considers predictors of the form $\sum_{i=1}^n \alpha_i k_i(x, x_i)$, with the regularization by the ℓ_1-norm of α instead of $\alpha^\top K \alpha$ for the regular single kernel learning problem.

Multivariate additive splines (MARS). This method of Friedman (1991) is the closest in spirit to the one presented in this paper: it builds in a forward greedy way multivariate piecewise polynomial expansions. Note however, that in MARS, a node is added only after one of its parents (and not all, like in HKL). We use the R package with standard hyperparameter settings.

Regression trees (CART). We consider regular decision trees for regression using the standard R implementation (Breiman et al., 1984) with standard hyperparameter settings.

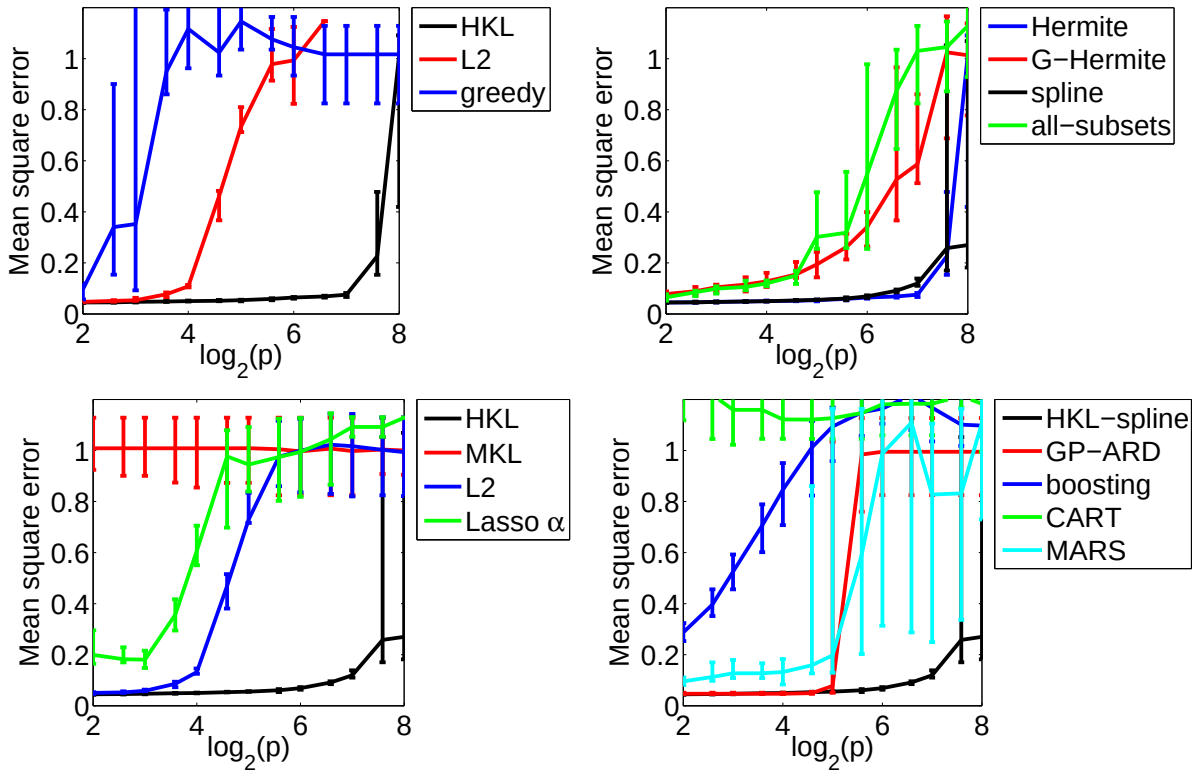

Figure 8: Comparison of non-linear regression methods (mean squared error vs. dimension of problem (in log scale). (Top left) comparison of greeedy, ℓ_2 and ℓ_1 (HKL) methods on the same Hermite kernel decomposition. (Top right) comparison of several kernel decompositions for HKL. (Bottom left) comparison with other kernel-based methods. (Bottom right) comparison with non-kernel-based methods.

Boosted regression trees (boosting). We use the R "gbm" package which implements the method of Friedman (2001).

Gaussian processes with automatic relevance determinations (GP-ARD). We use the code of Rasmussen and Williams (2006), which learns widths for each variable within a Gaussian kernel, using a Bayesian model selection criterion (i.e., without using cross-validation). Note that HKL, with the all-subset Gaussian decomposition, does not search explictly for A in the kernel $\exp(-(x-x')^\top A(x-x'))$, but instead considers a large set of particular values of A and finds a linear combination of the corresponding kernel.

6.2 Synthetic Examples

We generated synthetic data as follows: we generate a covariance matrix from a Wishart distribution of dimension p and with $2p$ degrees of freedom. It is then normalized to unit diagonal and n datapoints are then sampled i.i.d. from a Gaussian distribution with zero mean and this covariance matrix. We then consider the non-linear function $f(X) = \sum_{i=1}^{r}\sum_{j=i+1}^{r} X_j X_i$, which takes all cross products of the first r variables. The output Y is then equal to $f(X)$ plus some Gaussian noise with known signal-to-noise ratio.

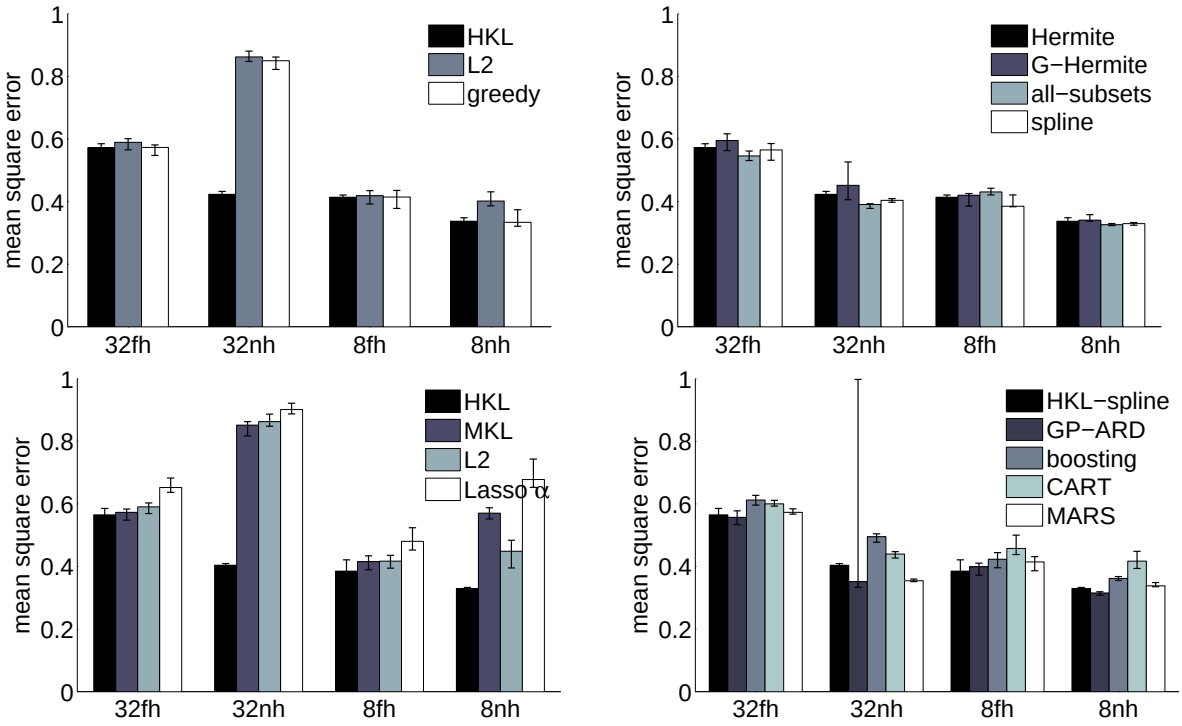

Figure 9: Comparison of non-linear regression methods (mean squared error vs. dimension of problem (in $\log$ scale). (Top left) comparison of greeedy, ℓ_2 and ℓ_1 (HKL) methods on the same Hermite kernel decomposition. (Top right) comparison of several kernel decompositions for HKL. (Bottom left) comparison with other kernel-based methods. (Bottom right) comparison with other non-kernel-based methods.

Results are reported in Figure 8. On the top left plot, we compare different strategies for linear regression, showing that in this constrained scenario where the generating model is sparse, ℓ_1-regularization based methods outperform other methods (forward selection and ridge regression). On the top right plot, we compare different kernel decompositions: as should be expected, the Hermite and spline decompositions (which contains exactly the generating polynomial) performs best. On the bottom left plot, we compare several kernel-based methods on the same spline kernel, showing that when sparsity is expected, using sparse methods is indeed advantageous. Finally, on the bottom right plot, we compare to non-kernel based methods, showing that ours is more robust to increasing input dimensions p. It is also worth noting the instabilities of the greedy methods such as MARS or "greedy", which sometimes makes wrong choices at the start of the procedure, leading to low performance.

6.3 UCI Datasets

We perform simulations on the "pumadyn" datasets from the UCI repository (Blake and Merz, 1998). These datasets are obtained from realistic simulations of the dynamics of a robot arm, and have different strengths of non-linearities (fh: fairly linear, high noise; nh: non-linear, high noise) and two numbers (8 and 32) of input variables.

Results are reported in Figure 9. On the top left plot, we compare different strategies for linear regression with $n = 1024$ observations: with moderately non-linear problems (32fh, 8fh), all performances are similar, while for non-linear problems (32nh, 8nh), HKL outperforms other methods (forward selection and ridge regression). On the top right plot, we compare different kernel decompositions: here, no decomposition includes the generating model, and therefore, none clearly outperforms the other ones. On the bottom left plot, we compare several kernel-based methods on the same spline kernel: it is interesting to note that for moderately linear problems, MKL performs well as expected, but not anymore for highly non-linear problems.

Finally, on the bottom right plot, we compare to non-kernel based methods: while boosting methods and CART are clearly performing worse, HKL, MARS and Gaussian processes perform better, with a significant advantage to MARS and Gaussian processes for the dataset "32nh". There are several explanations regarding the worse performance of HKL that could lead to interesting developments for improved performance: first, HKL relies on estimating a regularization parameter by cross-validation, while both MARS and GP-ARD rely on automatic model selection through frequentist or Bayesian procedures, and it is thus of clear interest to consider methods to automatically tune the regularization parameter for sparse methods such as HKL. Moreover, the problem is not really high-dimensional as n is much larger than p, and our regularized method has a certain amount of bias that the other methods don't have; this is a classical problem of ℓ_1-regularized problems, and this could be fixed by non-regularized estimation on the selected variables.

7 Conclusion

We have shown how to perform hierarchical multiple kernel learning (HKL) in polynomial time in the number of selected kernels. This framework may be applied to many positive definite kernels and we have focused on kernel decompositions for non-linear variable selection: in this setting, we can both select which variables should enter and the corresponding degrees of interaction complexity. We have proposed an active set algorithm as well a theoretical analysis that suggests that we can still perform *non-linear* variable selection from a number of variables which is exponential in the number of observations.

Our framework can be extended in multiple ways: first, this paper shows that trying to use ℓ_1-type penalties may be advantageous inside the feature space. That is, one may take the opposite directions than usual kernel-based methods and look inside the feature spaces with sparsity-inducing norms instead of building feature spaces of ever increasing dimensions. We are currently investigating applications to other kernels, such as the pyramid match kernels (Grauman and Darrell, 2007; Cuturi and Fukumizu, 2006), string kernels, and graph kernels (see, e.g., Shawe-Taylor and Cristianini, 2004). Moreover, theoretical and algorithmic connections with the recent work of Huang et al. (2009) on general structured sparsity and greedy methods could be made.

Moreover, we have considered in this paper a specific instance of block ℓ_1-norms with overlapping groups, i.e., groups organized in a hierarchy, but some of the techniques and frameworks presented here can be extended to more general overlapping structures (Jenatton et al., 2009), for DAGs or more general graphs; it would also be interesting to consider non discrete hierarchical structures with a partial order, such as positive definite matrices.

Finally, we hope to make connections with other uses of sparsity-inducing norms, in particular in signal processing, for compressed sensing (Baraniuk, 2007; Candès and Wakin, 2008), dictionary

learning (Olshausen and Field, 1997) and sparse principal component analysis (d'Aspremont et al., 2007).

A Proofs of Optimization Results

In this first appendix, we give proofs of all results related to the optimization problems.

A.1 Set of Weights for Trees

We prove that the set of weights ζ, i.e., Z, is itself convex when the DAG is a tree. We conjecture that the converse is true as well.

Proposition 9 *If V is a tree, the set $Z = \{\zeta(\eta) \in \mathbb{R}^V, \eta \in \mathbb{R}_+^V, \sum_{v\in V} d_v^2 \eta_v \leqslant 1\}$ is convex.*

Proof When the DAG is a tree (i.e., when each vertex has at most one parent and there is a single source r), then, we have for all v which is not the source of the DAG (i.e., for which there is exactly one parent), $\zeta_{\pi(v)}^{-1} - \zeta_v^{-1} = -\eta_v^{-1}$. This implies that the constraint $\eta \geqslant 0$ is equivalent to $\zeta_v \geqslant 0$ for all leaves v, and for all v which is not a source, $\zeta_{\pi(v)} \geqslant \zeta_v$, with equality possible only when they are both equal to zero.

Moreover, for the source r, $\zeta_r = \eta_r$. The final constraint $\sum_{v\in V} \eta_v d_v^2 \leqslant 1$, may then be written as $\sum_{v\neq r} d_v^2 \frac{1}{\zeta_v^{-1} - \zeta_{\pi(v)}^{-1}} + \zeta_r d_r^2 \leqslant 1$, that is, $\sum_{v\neq r} d_v^2 \left(\zeta_v + \frac{\zeta_v^2}{\zeta_{\pi(v)} - \zeta_v}\right) + \zeta_r d_r^2 \leqslant 1$, which is a convex constraint (Boyd and Vandenberghe, 2003). ■

A.2 Proof of Proposition 1

We introduce auxiliary variables $u_i = \langle f, \Phi(x_i)\rangle + b$ and consider the Lagrangian:

$$\mathcal{L}(u, f, b, \alpha) = \frac{1}{n}\sum_{i=1}^n \varphi_i(u_i) + \frac{\lambda}{2}\|f\|^2 + \lambda \sum_{i=1}^n \alpha_i(u_i - \langle f, \Phi(x_i)\rangle - b).$$

Minimizing with respect to the primal variable u leads to the term $-\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i)$; minimizing with respect to f leads to the term $-\frac{\lambda}{2}\alpha^\top K\alpha$ and to the expression of f as a function of α, and minimizing with respect to b leads to the constraint $1_n^\top \alpha = \sum_{i=1}^n \alpha_i = 0$.

A.3 Preliminary Propositions

We will use the following simple result, which implies that each component $\zeta_w(\eta)$ is a concave function of η (as the minimum of linear functions of η):

Lemma 10 *Let $a \in (\mathbb{R}_+^*)^m$. The minimum of $\sum_{j=1}^m a_j x_j^2$ subject to $x \geqslant 0$ and $\sum_{j=1}^m x_j = 1$ is equal to $\left(\sum_{j=1}^m a_i^{-1}\right)^{-1}$ and is attained at $x_i = a_i^{-1}\left(\sum_{j=1}^m a_i^{-1}\right)^{-1}$.*

Proof The result is a consequence of applying Cauchy-Schwartz inequality, applied to vectors with components $x_j a_j^{1/2}$ and $a_j^{-1/2}$. Note that when some of the a_j are equal to zero, then the minimum is zero, with optimal x_j being zero whenever $a_j \neq 0$. ■

The following proposition derives the dual of the problem in η, i.e., the dual of Eq. (17):

Proposition 11 *Let* $L = \{\kappa \in \mathbb{R}_+^{V\times V},\ \kappa_{\mathrm{A}(w)^c w} = 0$ *and* $\forall w \in V,\ \sum_{v\in\mathrm{A}(w)} \kappa_{vw} = 1\}$. *The following convex optimization problems are dual to each other, and there is no duality gap :*

$$\min_{\kappa\in L}\left\{\max_{v\in V} d_v^{-2} \sum_{w\in\mathrm{D}(v)} \kappa_{vw}^2 \alpha^\top \widetilde{K}_w \alpha\right\}, \tag{25}$$

$$\max_{\eta\in H} \sum_{w\in V} \zeta_w(\eta)\alpha^\top \widetilde{K}_w \alpha. \tag{26}$$

Proof We have the Lagrangian $\mathcal{L}(A,\kappa,\eta) = A + \sum_{v\in V}\eta_v\left(\sum_{w\in\mathrm{D}(v)}\kappa_{vw}^2\alpha^\top\widetilde{K}_w\alpha - Ad_v^2\right)$, with $\eta \geqslant 0$, which, using Lemma 10, can be minimized in closed form with respect to A, to obtain the constraints $\sum_{v\in V}\eta_v d_v^2 = 1$ and with respect to $\kappa \in L$. We thus get

$$\begin{aligned}\min_{\kappa\in L}\max_{v\in V} d_v^{-2}\sum_{w\in\mathrm{D}(v)}\kappa_{vw}^2\alpha^\top\widetilde{K}_w\alpha &= \max_\eta \alpha^\top\left(\sum_{w\in V}\left(\textstyle\sum_{v\in\mathrm{A}(w)}\eta_v^{-1}\right)^{-1}\widetilde{K}_w\right)\alpha,\\ &= \max_\eta \alpha^\top\left(\sum_{w\in V}\zeta_w(\eta)\widetilde{K}_w\right)\alpha.\end{aligned}$$

Given η, the optimal value for κ has a specific structure (using Lemma 10, for all $w \in V$): (a) if for all $v \in \mathrm{A}(w)$, $\eta_v > 0$, then $\kappa_{vw} = \zeta_w\eta_v^{-1}$ for all $v \in \mathrm{A}(w)$, (b) if there exists $v \in \mathrm{A}(w)$ such that $\eta_v = 0$, then for all $v \in \mathrm{A}(w)$ such that $\eta_v > 0$, we must have $\kappa_{vw} = 0$. ■

A.4 Proof of Proposition 3

We consider the following function of $\eta \in H$ and $\alpha \in \mathbb{R}^n$ (such that $1_n^\top\alpha = 0$):

$$F(\eta,\alpha) = -\frac{1}{n}\sum_{i=1}^n \psi_i(-n\lambda\alpha_i) - \frac{\lambda}{2}\alpha^\top\left(\sum_{w\in V}\zeta_w(\eta)\widetilde{K}_w\right)\alpha.$$

This function is convex in η (because of Lemma 10) and concave in α; standard arguments (e.g., primal and dual strict feasibilities) show that there is no duality gap to the variational problems:

$$\inf_{\eta\in H}\sup_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0} F(\eta,\alpha) = \sup_{\alpha\in\mathbb{R}^n,\ 1_n^\top\alpha=0}\inf_{\eta\in H} F(\eta,\alpha).$$

We can decompose the duality gap, given a pair (η, α) (with associated ζ, f and b) as:

$$
\begin{aligned}
& \sup_{\alpha' \in \mathbb{R}^n,\, 1_n^\top \alpha' = 0} F(\eta, \alpha') - \inf_{\eta' \in H} F(\eta', \alpha) \\
= \;& \min_{f,b} \left\{ \frac{1}{n} \sum_{i=1}^n \varphi_i \Big(\sum_{v \in V} \langle f_v, \Phi_v(x_i) \rangle + b \Big) + \frac{\lambda}{2} \sum_{w \in V} \zeta_w(\eta)^{-1} \|f_w\|^2 \right\} - \inf_{\eta' \in H} F(\eta', \alpha), \\
\leqslant \;& \frac{1}{n} \sum_{i=1}^n \varphi_i \Big(\sum_{w \in V} \zeta_w(\eta) (K_w \alpha)_i + b \Big) + \frac{\lambda}{2} \sum_{w \in V} \zeta_w \alpha^\top \widetilde{K}_w \alpha + \frac{1}{n} \sum_{i=1}^n \psi_i(-n\lambda\alpha_i) \\
& + \sup_{\eta' \in H} \frac{\lambda}{2} \alpha^\top \sum_{w \in V} \zeta_w(\eta') \alpha, \\
= \;& \frac{1}{n} \sum_{i=1}^n \varphi_i \Big(\sum_{w \in V} \zeta_w(\eta) (K_w \alpha)_i + b \Big) + \frac{1}{n} \sum_{i=1}^n \psi_i(-n\lambda\alpha_i) + \lambda \sum_{w \in V} \zeta_w(\eta) \alpha^\top \widetilde{K}_w \alpha \\
& + \frac{\lambda}{2} \Big[\sup_{\eta' \in H} \sum_{w \in V} \zeta_w(\eta') \alpha^\top K_w \alpha - \sum_{w \in V} \zeta_w(\eta) \alpha^\top \widetilde{K}_w \alpha \Big], \\
= \;& \mathrm{gap}_{\mathrm{kernel}} \Big(\sum_{w \in V} \zeta_w(\eta) \widetilde{K}_w, \alpha \Big) + \frac{\lambda}{2} \mathrm{gap}_{\mathrm{weights}} \Big((\alpha^\top \widetilde{K}_w \alpha)_{w \in V}, \eta \Big).
\end{aligned}
$$

We thus get the desired upper bound from which Proposition 3 follows, as well as the upper bound on the duality gap in Eq. (18).

A.5 Proof of Propositions 4 and 5

We assume that we know the optimal solution of a truncated problem where the entire set of decendants of some nodes have been removed. We let denote W the hull of the set of active variables. We now consider necessary conditions and sufficient conditions for this solution to be optimal with respect to the full problem. This will lead to Propositions 4 and 5.

We first use Proposition 11, to get a set of κ_{vw} for $(v, w) \in W$ for the reduced problem; the goal here is to get necessary conditions by relaxing the dual problem in Eq. (25), defining $\kappa \in L$ and find an approximate solution, while for the sufficient condition, any candidate leads to a sufficient condition. It turns out that we will use the solution of the relaxed solution required for the necessary condition for the sufficient condition.

Necessary condition. If we assume that all variables in W are active and the reduced set is optimal for the full problem, then any optimal $\kappa \in L$ must be such that $\kappa_{vw} = 0$ if $v \in W$ and $w \in W^c$, and we must have $\kappa_{vw} = \zeta_w \eta_v^{-1}$ for $v \in W$ and $w \in \mathrm{D}(v) \cap W$ (otherwise, η_W cannot be optimal for the reduced problem, as detailed in the proof of Proposition 11). We then let free κ_{vw} for v, w in W^c. Our goal is to find good candidates for those free dual parameters.

We can lowerbound the sums by maxima:

$$
\max_{v \in V \cap W^c} d_v^{-2} \sum_{w \in \mathrm{D}(v)} \kappa_{vw}^2 \alpha^\top \widetilde{K}_w \alpha \geqslant \max_{v \in V \cap W^c} d_v^{-2} \max_{w \in \mathrm{D}(v)} \kappa_{vw}^2 \alpha^\top \widetilde{K}_w \alpha,
$$

which can be minimized in closed form with respect to κ leading to $\kappa_{vw} = d_v \left(\sum_{v' \in A(w) \cap W^c} d_{v'}\right)^{-1}$ and, owing to Proposition 11 to the following lower bound for $\max_{\eta \in H} \sum_{w \in V} \zeta_w(\eta) \alpha^\top \widetilde{K}_w \alpha$:

$$\max\left\{\delta^2, \max_{w \in W^c} \frac{\alpha^\top \widetilde{K}_w \alpha}{(\sum_{v \in A(w) \cap W^c} d_v)^2}\right\} \geqslant \max\left\{\delta^2, \max_{w \in \mathrm{sources}(W^c)} \frac{\alpha^\top \widetilde{K}_w \alpha}{(\sum_{v \in A(w) \cap W^c} d_v)^2}\right\}, \quad (27)$$

where $\delta^2 = \sum_{w \in W} \zeta_w(\eta_W) \alpha^\top \widetilde{K}_w \alpha = \Omega(f)^2$. If the reduced solution is optimal we must have this lower bound smaller than δ^2, which leads to Eq. (20). Note that this necessary condition may also be obtained by considering the addition (alone) of any of the sources $w \in \mathrm{sources}(W^c)$ and checking that they would not enter the active set.

Sufficient condition. For sufficient conditions, we simply take the previous value obtained before for κ, which leads to the following upperbound for $\max_{\eta \in H} \sum_{w \in V} \zeta_w(\eta) \alpha^\top \widetilde{K}_w \alpha$:

$$\max\left\{\delta^2, \max_{t \in W^c} \sum_{w \in \mathrm{D}(t)} \frac{\alpha^\top \widetilde{K}_w \alpha}{(\sum_{v \in A(w) \cap W^c} d_v)^2}\right\} = \max\left\{\delta^2, \max_{t \in \mathrm{sources}(W^c)} \sum_{w \in \mathrm{D}(t)} \frac{\alpha^\top \widetilde{K}_w \alpha}{(\sum_{v \in A(w) \cap W^c} d_v)^2}\right\},$$

because for all $v \in W^c$, there exists $t \in \mathrm{sources}(W^c)$ such that $v \in \mathrm{D}(t)$. We have moreover for all $t \in W^c$,

$$\sum_{v \in A(w) \cap W^c} d_v \geqslant \sum_{v \in A(w) \cap D(t)} d_v,$$

leading to the upper bound: $A = \max\left\{\delta^2, \max_{t \in \mathrm{sources}(W^c)} \sum_{w \in D(t)} \frac{\alpha^\top \widetilde{K}_w \alpha}{(\sum_{v \in A(w) \cap D(t)} d_v)^2}\right\}$. The gap in Eq. (18) is thus less than $\lambda/2(A - \delta^2)$, which leads to the desired result.

A.6 Optimality Conditions for the Primal Formulation

We now derive optimality conditions for the primal problem in Eq. (13), when the loss functions φ_i are differentiable, which we will need in Appendix B, that is:

$$\min_{f \in \mathcal{F},\, b \in \mathbb{R}} L(f, b) + \frac{\lambda}{2} \Omega(f)^2,$$

where $L(f, b)$ is the differentiable loss function. Following Bach (2008a) and Proposition 2, the solution may be found by solving a finite-dimensional problem, and thus usual notions of calculus may be used.

Let $f \in \mathcal{F} = \prod_{v \in V} \mathcal{F}_v$ and $b \in \mathbb{R}$, where $f \neq 0$, with W being the hull of the active functions (or groups). The directional derivative in the direction $(\Delta, \tau) \in \mathcal{F}^V \times \mathbb{R}$ is equal to

$$\langle \nabla_f L(f, b), \Delta \rangle + \nabla_b L(f, b) \tau + \lambda \Omega(f) \left(\sum_{v \in W} d_v \left\langle \frac{f_{\mathrm{D}(v)}}{\|f_{\mathrm{D}(v)}\|}, \Delta_v \right\rangle + \sum_{v \in W^c} d_v \|\Delta_{\mathrm{D}(v)}\| \right),$$

and thus (f, b) if optimal if and ony if $\nabla_b L(f, b) = 0$ (i.e., b is an optimal constant term) and if, with $\delta = \Omega(f)$:

$$\forall w \in W,\ \nabla_{f_w} L(f, b) + \lambda \delta \left(\sum_{v \in \mathrm{A}(w)} \frac{d_v}{\|f_{\mathrm{D}(v)}\|} \right) f_w = 0, \quad (28)$$

$$\text{and } \forall \Delta_{W^c} \in \mathbb{R}^{W^c}, \ \sum_{w \in W^c} \langle \nabla_{f_w} L(f,b), \Delta_w \rangle + \lambda \delta \Big(\sum_{v \in W^c} d_v \|\Delta_{\mathrm{D}(v)}\| \Big) \geqslant 0. \tag{29}$$

We can now define for $K \subset V$, $\Omega_K(f_K) = \sum_{v \in K} d_v \|f_{\mathrm{D}(v) \cap K}\|$, the norm reduced to the functions in K and Ω_K^* its dual norm (Boyd and Vandenberghe, 2003; Rockafellar, 1970). The last equation may be rewritten: $\Omega^*_{W^c}(\nabla_{f_W} L(f,b)) \leqslant \lambda\delta$. Note that when regularizing by $\lambda\Omega(f) = \lambda \sum_{v \in V} d_v \|f_{\mathrm{D}(v)}\|$ instead of $\frac{\lambda}{2} \Big(\sum_{v \in V} d_v \|f_{\mathrm{D}(v)}\| \Big)^2$, we have the same optimality condition with $\delta = 1$.

B Proof of Theorem 7

In this appendix, we provide the proof of Theorem 7 with several intermediate results. Following usual proof techniques from the Lasso literature, we will consider the optimization reduced to kernels/variables in $\mathbf{W}$, and (a) show that the hull of the selected variables is indeed the hull of $\mathbf{W}$ (i.e., itself because we have assumed in (**A0**) that $\mathbf{W}$ is equal to its hull) with high probability, and (b) show that when the reduced solution is extended to $\mathbf{W}^c$ with zeros, we have the optimal global solution of the problem with high probability. The main difficulties are to use bounds on the dual norms of our structured norms, and to deal with the infinite-dimensional group structure within a non-asymptotic analysis, which we deal with new concentration inequalities (Appendices B.2, B.3 and B.4).

B.1 Notations

Let $\hat{\mu}_v = \frac{1}{n} \sum_{i=1}^n \Phi_v(x_i) \in \mathcal{F}_v$ be the empirical mean and $\boldsymbol{\mu}_v = \mathbb{E}\Phi_v(X) \in \mathcal{F}_v$ the population mean of $\Phi_v(X)$ and $\widehat{\Sigma}_{vw} = \frac{1}{n} \sum_{i=1}^n (\Phi_v(x_i) - \hat{\mu}_v) \otimes (\Phi_w(x_i) - \hat{\mu}_w)$ be the empirical cross-covariance operator from $\mathcal{F}_w$ to $\mathcal{F}_v$ and $q_v = \frac{1}{n} \sum_{i=1}^n \varepsilon_i (\Phi_v(x_i) - \hat{\mu}_v) \in \mathcal{F}_v$ for $v, w \in V$, where $\varepsilon_i = y_i - \sum_{w \in \mathbf{W}} \mathbf{f}_w(x_i) - \mathbf{b}$ is the i.i.d. Gaussian noise with mean zero and variance $\boldsymbol{\sigma}^2$. By assumption (**A2**), we have $\operatorname{tr} \boldsymbol{\Sigma}_{vv} \leqslant 1$ and $\operatorname{tr} \widehat{\Sigma}_{vv} \leqslant 1$ for all $v \in V$, which implies that $\lambda_{\max}(\boldsymbol{\Sigma}_{\mathbf{WW}}) \leqslant |\mathbf{W}|$ and $\lambda_{\max}(\widehat{\Sigma}_{\mathbf{WW}}) \leqslant |\mathbf{W}|$.

All norms on vectors in Euclidean or Hilbertian spaces are always the Euclidean or Hilbertian norms of the space the vector belongs to (which can always be inferred from context). However, we consider several norms on self-adjoint operators between Hilbert spaces. All our covariance operators are *compact* and can thus be diagonalized in an Hilbertian basis, with a sequence of eigenvalues that tends to zero (see, e.g., Brezis, 1980; Berlinet and Thomas-Agnan, 2003; Conway, 1997). The usual operator norm of a self-adjoint operator A is the eigenvalue of largest magnitude of A and is denoted by $\|A\|_{\mathrm{op}}$; the Hilbert-Schmidt norm is the ℓ_2-norm of eigenvalues, and is denoted by $\|A\|_{\mathrm{HS}}$, and is equal to the Frobenius norm in finite dimensions. Finally, the trace norm is equal to the ℓ_1-norm of eigenvalues, and is denoted by $\|A\|_{\mathrm{tr}}$. In Section B.3, we provide novel non asymptotic results on the convergence of empirical covariance operators to the population covariance operators.

B.2 Hoeffding's Inequality in Hilbert Spaces

In this section, we prove the following proposition, which will be useful throughout this appendix:

Proposition 12 *Let $X_1, \ldots, X_n$ be i.i.d. zero-mean random observations in the Hilbert space $\mathcal{H}$, such that for all i, $\|X_i\| \leqslant 1$ almost surely. Then, we have:*

$$\mathbb{P}\Big(\Big\|\frac{1}{n}\sum_{i=1}^n X_i\Big\| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{nt^2}{8}\Big). \tag{30}$$

Proof We denote $Z = \big\|\frac{1}{n}\sum_{i=1}^n X_i\big\|$. If all X_i are held fixed but one, then Z may only change by $\frac{2}{n}$. Thus, from Mc Diarmid's inequality (see, e.g., Massart, 2003, Theorem 5.1, page 148), we have, for all $t \geqslant 0$:

$$\mathbb{P}(Z - \mathbb{E}Z \geqslant t) \leqslant \exp(-nt^2/2).$$

Moreover, using the Hilbertian structure of $\mathcal{H}$:

$$\mathbb{E}Z \leqslant (\mathbb{E}Z^2)^{1/2} = \left(\frac{1}{n^2}\sum_{i,j=1}^n \mathbb{E}\langle X_i, X_j\rangle\right)^{1/2} = n^{-1/2}(\mathbb{E}\|X_1\|^2)^{1/2} \leqslant n^{-1/2}.$$

This leads to $\mathbb{P}(Z \geqslant n^{-1/2}t + n^{-1/2}) \leqslant \exp(-t^2/2)$ for all $t \geqslant 0$, i.e., for all $t \geqslant 1$, $\mathbb{P}(Z \geqslant tn^{-1/2}) \leqslant \exp(-(t-1)^2/2)$. If $t \geqslant 2$, then $(t-1)^2 \geqslant t^2/4$, and thus $\mathbb{P}(Z \geqslant tn^{-1/2}) \leqslant \exp(-t^2/8) \leqslant 2\exp(-nt^2/8)$. For $t \leqslant 2$, then the right hand side is greater than $2\exp(-1/2) > 1$, and the bound in Eq. (30) is trivial. ■

B.3 Concentration Inequalities for Covariance Operators

We prove the following general proposition of concentration of empirical covariance operators for the Hilbert-Schmidt norm:

Proposition 13 *Let $X_1, \ldots, X_n$ be i.i.d. random observations in a measurable space $\mathcal{X}$, equipped with a reproducing kernel Hilbert space $\mathcal{F}$ with kernel k, such that $k(X_i, X_i) \leqslant 1$ almost surely. Let $\boldsymbol{\Sigma}$ and $\widehat{\Sigma}$ be the population and empirical covariance operators. We have, for all $x \geqslant 0$:*

$$\mathbb{P}(\|\boldsymbol{\Sigma} - \widehat{\Sigma}\|_{\mathrm{HS}} \geqslant xn^{-1/2}) \leqslant 4\exp\Big(-\frac{x^2}{32}\Big).$$

Proof We first concentrate the mean, using Proposition 12, since the data is universally bounded by 1:

$$\mathbb{P}(\|\hat{\mu} - \boldsymbol{\mu}\| \geqslant t) \leqslant 2\exp\Big(-\frac{nt^2}{8}\Big).$$

The random variables $(\Phi(X_i) - \boldsymbol{\mu}) \otimes (\Phi(X_i) - \boldsymbol{\mu})$ are uniformly bounded by 1 in the Hilbert space of self-adjoint operators, equipped with the Hilbert-Schmidt norm. Thus, using Proposition 12, we get

$$\mathbb{P}\Big(\Big\|\boldsymbol{\Sigma} - \frac{1}{n}\sum_{i=1}^n (\Phi(X_i) - \boldsymbol{\mu}) \otimes (\Phi(X_i) - \boldsymbol{\mu})\Big\|_{\mathrm{HS}} \geqslant x\Big) \leqslant 2\exp\Big(-\frac{nx^2}{8}\Big).$$

Thus, since $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (\Phi(X_i)-\boldsymbol{\mu})\otimes(\Phi(X_i)-\boldsymbol{\mu}) + (\boldsymbol{\mu}-\hat{\mu})\otimes(\boldsymbol{\mu}-\hat{\mu})$, and $\|(\boldsymbol{\mu}-\hat{\mu})\otimes(\boldsymbol{\mu}-\hat{\mu})\|_{\mathrm{HS}} = \|\boldsymbol{\mu}-\hat{\mu}\|^2$, we get:

$$\mathbb{P}(\|\boldsymbol{\Sigma}-\widehat{\Sigma}\|_{\mathrm{HS}} \geqslant x) \leqslant 2\exp\Big(-\frac{nx^2}{32}\Big) + 2\exp\Big(-\frac{nx}{16}\Big) \leqslant 4\exp\Big(-\frac{nx^2}{32}\Big),$$

as long as $x \leqslant 2$. When $x > 2$, the bound is trivial because $\|\boldsymbol{\Sigma}-\widehat{\Sigma}\|_{\mathrm{HS}} \geqslant x$ occurs with probability zero. ■

We now prove the following general proposition of concentration of empirical covariance operators for the *trace norm*:

Proposition 14 *Let $X_1,\ldots,X_n$ be i.i.d. random observations in a measurable space $\mathcal{X}$, equipped with a reproducing kernel Hilbert space $\mathcal{F}$ with kernel k, such that $k(X_i,X_i) \leqslant 1$ almost surely. Let $\boldsymbol{\Sigma}$ and $\widehat{\Sigma}$ the population and empirical covariance operators. Assume that the eigenvalues of $\boldsymbol{\Sigma}$ are root-summable with sum of square roots of eigenvalues equal to $C_{1/2}$. We have, if $x \geqslant 4C_{1/2}$:*

$$\mathbb{P}(\|\boldsymbol{\Sigma}-\widehat{\Sigma}\|_{\mathrm{tr}} \geqslant xn^{-1/2}) \leqslant 3\exp\Big(-\frac{x^2}{32}\Big).$$

Proof It is shown by Harchaoui et al. (2008) that

$$\mathbb{E}\|\boldsymbol{\Sigma}-\widehat{\Sigma}\|_{\mathrm{tr}} \leqslant C_{1/2}n^{-1/2}.$$

Thus, following the same reasoning as in the proof of Proposition 12, we get

$$\mathbb{P}\Big(\Big\|\boldsymbol{\Sigma} - \frac{1}{n}\sum_{i=1}^n (\Phi(X_i)-\boldsymbol{\mu})\otimes(\Phi(X_i)-\boldsymbol{\mu})\Big\|_{\mathrm{tr}} \geqslant (C_{1/2}+t)n^{-1/2}\Big) \leqslant \exp(-t^2/2),$$

and thus if $t \geqslant 2C_{1/2}$, we have:

$$\mathbb{P}\Big(\Big\|\boldsymbol{\Sigma} - \frac{1}{n}\sum_{i=1}^n (\Phi(X_i)-\boldsymbol{\mu})\otimes(\Phi(X_i)-\boldsymbol{\mu})\Big\|_{\mathrm{tr}} \geqslant tn^{-1/2}\Big) \leqslant \exp(-t^2/8).$$

We thus get, for $x \geqslant 4C_{1/2}$,

$$\mathbb{P}(\|\boldsymbol{\Sigma}-\widehat{\Sigma}\|_{\mathrm{tr}} \geqslant xn^{-1/2}) \leqslant \exp(-x^2/32) + 2\exp\Big(-\frac{xn^{+1/2}}{16}\Big) \leqslant 3\exp(-x^2/32),$$

as long as $xn^{-1/2} \leqslant 2$. If this is not true, the bound to be proved is trivial. ■

B.4 Concentration Inequality for Least-squares Problems

In this section, we prove a concentration result that can be applied to several problems involving least-squares and covariance operators (Harchaoui et al., 2008; Fukumizu et al., 2007; Bach, 2008a):

Proposition 15 *Let $X_1,\ldots,X_n$ be i.i.d. random observations in a measurable space $\mathcal{X}$, equipped with a reproducing kernel Hilbert space $\mathcal{F}$ with kernel k, such that $k(X_i,X_i) \leqslant 1$ almost surely. Let $\mathbf{\Sigma}$ and $\widehat{\Sigma}$ the population and empirical covariance operators. Assume that the eigenvalues of $\mathbf{\Sigma}$ are root-summable with sum of square roots of eigenvalues equal to $C_{1/2}$. Let ε be an independent Gaussian vector with zero mean and covariance matrix $\boldsymbol{\sigma}^2\mathrm{I}$. Define $q = \frac{1}{n}\sum_{i=1}^n \varepsilon_i(\Phi(X_i) - \hat{\mu})$. We have, for all $t \geqslant \left(4\boldsymbol{\sigma}^2 n^{-1}\left[\lambda^{-1/2}C_{1/2} + \|\widehat{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{tr}}\lambda^{-1}\right]\right)^{1/2}$:*

$$\mathbb{P}(\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1/2}q\| \geqslant t|X) \leqslant \exp(-nt^2/2\boldsymbol{\sigma}^2)$$

Proof Given the input variables, $\|(\widehat{\Sigma}+\lambda\mathrm{I})^{-1/2}q\|$ is a Lipschitz-continuous function of the i.i.d. noise vector ε, with Lipschitz constant $n^{-1/2}$. Moreover, we have

$$\mathbb{E}\left(\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1/2}q\||X\right) \leqslant \mathbb{E}\left(\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1/2}q\|^2|X\right)^{1/2} = \boldsymbol{\sigma}n^{-1/2}\left(\mathrm{tr}\,\widehat{\Sigma}(\widehat{\Sigma} + \lambda\mathrm{I})^{-1}\right)^{1/2}.$$

We now follow Harchaoui et al. (2008) for bounding the empirical degrees of freedom:

$$\begin{aligned}
& \mathrm{tr}\,\widehat{\Sigma}(\widehat{\Sigma} + \lambda\mathrm{I})^{-1} - \mathrm{tr}\,\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda\mathrm{I})^{-1} \\
= \;& \lambda\,\mathrm{tr}(\mathbf{\Sigma} + \lambda\mathrm{I})^{-1}(\widehat{\Sigma} - \mathbf{\Sigma})(\widehat{\Sigma} + \lambda\mathrm{I})^{-1} \\
\leqslant \;& \lambda\|\widehat{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{tr}}\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1}\|_{\mathrm{op}}\|(\mathbf{\Sigma} + \lambda\mathrm{I})^{-1}\|_{\mathrm{op}} \leqslant \lambda^{-1}\|\widehat{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{tr}}.
\end{aligned}$$

Moreover, we have: $\mathrm{tr}\,\mathbf{\Sigma}(\mathbf{\Sigma} + \lambda\mathrm{I})^{-1} \leqslant \lambda^{-1/2}C_{1/2}$. This leads to:

$$\mathbb{E}\left(\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1/2}q\||X\right)^2 \leqslant \boldsymbol{\sigma}^2 n^{-1}\left[\lambda^{-1/2}C_{1/2} + \|\widehat{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{tr}}\lambda^{-1}\right].$$

The final bound is obtained from concentration of Lipschitz-continuous functions of Gaussian variables (Massart, 2003):

$$\mathbb{P}(\|(\widehat{\Sigma} + \lambda\mathrm{I})^{-1/2}q\| \geqslant t|X) \leqslant \exp(-nt^2/2\boldsymbol{\sigma}^2)$$

as soon as $t^2 \geqslant 4\boldsymbol{\sigma}^2 n^{-1}\left[\lambda^{-1/2}C_{1/2} + \|\widehat{\Sigma} - \mathbf{\Sigma}\|_{\mathrm{tr}}\lambda^{-1}\right]$. ■

B.5 Concentration Inequality for Irrelevant Variables

In this section, we upperbound, using Gaussian concentration inequalities (Massart, 2003), the tail-probability

$$\mathbb{P}(\Omega^*_{\mathbf{W}^c}[z] \geqslant t),$$

where $z = -q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{W}\mathbf{W}} + D)^{-1}q_{\mathbf{W}}$, for a given deterministic nonnegative diagonal matrix D. The vector z may be expressed as weighted sum of the components of the Gaussian vector ε. In addition, $\Omega^*_{\mathbf{W}^c}[g_{\mathbf{W}^c}]$ is upperbounded by $\max_{w\in\mathbf{W}^c}\|g_w\|d_w^{-1} \leqslant d_r^{-1}\max_{w\in\mathbf{W}^c}\|g_w\|$. Thus by concentration of Lipschitz-continuous functions of multivariate standard random variables (we have a $d_r^{-1}n^{-1/2}$-Lipschitz function of ε), we have (Massart, 2003):

$$\mathbb{P}[\Omega^*_{\mathbf{W}^c}[z] \geqslant t + \mathbb{E}(\Omega^*_{\mathbf{W}^c}[z]|x)|x] \leqslant \exp\left(-\frac{nt^2d_r^2}{2\boldsymbol{\sigma}^2}\right).$$

For all $w \in \mathbf{W}^c$, given $(x_1, \ldots, x_n)$, $n^{1/2}\boldsymbol{\sigma}^{-1} z_w \in \mathcal{F}_w$ is normally distributed with covariance operator which has largest eigenvalue less than one. We now decompose $\mathbf{W}^c$ by values of d_w: by assumption, d_w may take value d_r or a power of β (we let denote $\mathcal{D}$ the set of values of d_w, $w \in V$). We get (where x denotes all input observations):

$$\begin{aligned}
n^{1/2}\boldsymbol{\sigma}^{-1}\mathbb{E}\left(\max_{w\in\mathbf{W}^c}\frac{\|z_w\|}{d_w}\Big|x\right) &\leqslant n^{1/2}\boldsymbol{\sigma}^{-1}\sum_{d\in\mathcal{D}}\mathbb{E}\left(\max_{w\in\mathbf{W}^c,\ d_w=d}\frac{\|z_w\|}{d}\Big|x\right)\\
&\leqslant \sum_{d\in\mathcal{D}}\frac{2}{d}\log(2|\{w\in\mathbf{W}^c, d_w=d\}|)^{1/2}\\
&\leqslant \sum_{d\in\mathcal{D}}\frac{2}{d}\log(2|\{w\in V, d_w=d\}|)^{1/2}\\
&\leqslant d_r^{-1}\sum_{k\geqslant 0}\frac{2}{\beta^k}\log(2|\{w\in V,\ \mathrm{depth}(w)=k\}|)^{1/2}\\
&\leqslant d_r^{-1}\sum_{k\geqslant 0}\frac{2}{\beta^k}\log(2|\mathrm{depth}^{-1}(k)|)^{1/2} = d_r^{-1}A.
\end{aligned}$$

We thus get $\mathbb{P}\left[\Omega^*_{\mathbf{W}^c}[z] \geqslant \frac{\boldsymbol{\sigma}(t+A)}{d_r n^{1/2}}\Big|x\right] \leqslant \exp\left(-\frac{t^2}{2}\right)$, and if we use $t \geqslant 2A$, we get

$$\mathbb{P}\Big(\Omega^*_{\mathbf{W}^c}[Q] \geqslant \frac{\boldsymbol{\sigma} t}{d_r n^{1/2}}\Big|x\Big) \leqslant \exp\left(-\frac{t^2}{8}\right). \tag{31}$$

Note that we have used the expectation of the maximum of q norms of Gaussian vectors is less than $2(\log(2q))^{1/2}$ times the maximum of the expectation of the norms.

Upper bound on A. The cardinal of $\mathrm{depth}^{-1}(k)$ is less than $\mathrm{num}(V)\deg(V)^k$, thus, since $\beta > 1$,

$$\begin{aligned}
A &= \sum_{k\geqslant 0}\frac{2}{\beta^k}\log(2|\mathrm{depth}^{-1}(k)|)^{1/2}\\
&\leqslant \sum_{k\geqslant 0}\frac{2}{\beta^k}[(\log(2\mathrm{num}(V))^{1/2} + (k\log\deg(V))^{1/2}]\\
&\leqslant \frac{2}{1-\beta^{-1}}(\log(2\mathrm{num}(V))^{1/2} + (\log\deg(V))^{1/2}2\sum_{k\geqslant 0}\beta^{-k}k^{1/2}.
\end{aligned}$$

Moreover, we have, by splitting the sum at $(2\log\beta)^{-1}$, and using the fact that after the split, the function $x \mapsto \beta^{-x}x^{1/2}$ is decreasing:

$$\begin{aligned}
2\sum_{k\geqslant 0}\beta^{-k}k^{1/2} &\leqslant 2\sum_{k\geqslant 1}\beta^{-k}k^{1/2} \leqslant 2\sum_{k=1}^{(2\log\beta)^{-1}}\beta^{-k}k^{1/2} + 2\sum_{k=(2\log\beta)^{-1}}^{\infty}\beta^{-k}k^{1/2},\\
&\leqslant \frac{2}{(2\log\beta)^{3/2}} + 2\int_0^{+\infty}\beta^{-x}x^{1/2}dx,\\
&\leqslant \frac{2}{(2\log\beta)^{3/2}} + 2(\log\beta)^{-3/2}\int_0^{+\infty}e^{-x}x^{1/2}dx,\\
&\leqslant \frac{1}{(\log\beta)^{3/2}}\left(1+\Gamma(3/2)\right) \leqslant \frac{2}{(\log\beta)^{3/2}},\ \text{where } \Gamma(\cdot) \text{ is the Gamma function.}
\end{aligned}$$

This leads to $A \leqslant \frac{2}{1-\beta^{-1}}(\log(2\mathrm{num}(V))^{1/2} + (\log \deg(V))^{1/2}\frac{2}{(\log \beta)^{3/2}}$ and the expression for $\gamma(V)$ in Theorem 7.

B.6 Error of the Reduced Solution

We have the following loss function (optimized with respect to the constant term $b \in \mathbb{R}$)

$$L(f) = \frac{1}{2}\langle f - \mathbf{f}, \widehat{\Sigma}(f - \mathbf{f})\rangle - \langle q, f - \mathbf{f}\rangle.$$

Following Bach (2008a) and Nardi and Rinaldo (2008), we consider the reduced problem on $\mathbf{W}$, $\min_{f \in \mathcal{F},\, f_{\mathbf{W}^c}=0} L(f) + \lambda\Omega_{\mathbf{W}}(f_{\mathbf{W}})$, with non unique solution $\hat{f}$ (since $\widehat{\Sigma}_{\mathbf{WW}}$ is not invertible in general). The goal here is to show that $\hat{f}$ and $\mathbf{f}$ are close enough so that for all $w \in \mathbf{W}$, $\hat{f}_{\mathrm{D}(w)} \neq 0$; this will implies that the hull of the active set of $\hat{f}$ is indeed $\mathbf{W}$.

As opposed to the Lasso case, we also need to consider $\tilde{f}_{\mathbf{W}}$ the minimum of $f_{\mathbf{W}} \mapsto L(f_{\mathbf{W}}) + \frac{\lambda}{2}\sum_{v \in \mathbf{W}} \frac{\|f_w\|^2}{\boldsymbol{\zeta}_w}$, which corresponds to the local quadratic approximation of the norm around $\mathbf{f}_{\mathbf{W}}$, where

$$\boldsymbol{\zeta}_w^{-1} = \zeta_w(\mathbf{f}_{\mathbf{W}})^{-1} = \Omega(\mathbf{f}) \sum_{v \in \mathrm{A}(w)} \frac{d_v}{\|\mathbf{f}_{\mathrm{D}(v)}\|}.$$

Moreover, we consider the corresponding noiseless version $\tilde{\mathbf{f}}_{\mathbf{W}}$ of $\tilde{f}_{\mathbf{W}}$ (the solution for $\varepsilon = 0$). We will compute error bounds $\|\tilde{\mathbf{f}}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\|$, $\|\tilde{f}_{\mathbf{W}} - \tilde{\mathbf{f}}_{\mathbf{W}}\|$ and $\|\tilde{f}_{\mathbf{W}} - \hat{f}_{\mathbf{W}}\|$, which will provide an upper bound on $\|\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\|$ (see Proposition 19). In particular, once we have $\|\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \nu/2$, then we must have $\|\hat{f}_{\mathrm{D}(w)}\| > 0$ for all $w \in \mathbf{W}$ and thus the hull of selected kernels is indeed $\mathbf{W}$.

Lemma 16 *We have:*

$$\|\tilde{\mathbf{f}}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \left(\lambda + \|\widehat{\Sigma}_{\mathbf{WW}} - \boldsymbol{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}} d_r^{-2}\right) \frac{\Omega(\mathbf{f})^2 |\mathbf{W}|^{1/2}}{\kappa\nu}. \tag{32}$$

Proof The function $\tilde{\mathbf{f}}$ is defined as, with $D = \mathrm{Diag}(\boldsymbol{\zeta}_w^{-1}\mathrm{I})$,

$$\tilde{\mathbf{f}}_{\mathbf{W}} = (\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1}\widehat{\Sigma}_{\mathbf{WW}}\mathbf{f}_{\mathbf{W}} = \mathbf{f}_{\mathbf{W}} - \lambda(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}}.$$

Thus, we have

$$\begin{aligned}\|\tilde{\mathbf{f}}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \lambda \left\|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1}(\widehat{\Sigma}_{\mathbf{WW}} - \boldsymbol{\Sigma}_{\mathbf{WW}})(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}}\right\| \\ + \lambda \left\|(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}}\right\|.\end{aligned}$$

We can now upper bound $\left\|(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}}\right\| \leqslant \|\mathbf{h}_{\mathbf{W}}\|\kappa^{-1}\|D\|_{\mathrm{op}} \leqslant |\mathbf{W}|^{1/2}\kappa^{-1}\Omega(\mathbf{f})^2\nu^{-2}$.

$$\begin{aligned}\|\tilde{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| &\leqslant \left(\lambda + \|\widehat{\Sigma}_{\mathbf{WW}} - \boldsymbol{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}}\|D^{-1}\|_{\mathrm{op}}\right)\left\|(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}}\right\| \\ &\leqslant \left(\lambda + \|\widehat{\Sigma}_{\mathbf{WW}} - \boldsymbol{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}} d_r^{-2}\right)\frac{\Omega(\mathbf{f})^2}{\nu^2}|\mathbf{W}|^{1/2}\kappa^{-1}.\end{aligned}$$

We have used moreover the following identities:

$$\boldsymbol{\zeta}_w^{-1} \geqslant d_r^2 \quad \text{and} \quad \boldsymbol{\zeta}_w^{-1} = \Omega(\mathbf{f}) \sum_{v \in \mathrm{A}(w)} \frac{d_v}{\|\mathbf{f}_{\mathrm{D}(v)}\|} \leqslant \frac{\Omega(\mathbf{f})^2}{\nu^2},$$

which leads to $\|D^{-1}\|_{\rm op} \leqslant d_r^{-2}$ and $\|D\|_{\rm op} \leqslant \Omega(\mathbf{f})^2\nu^{-2}$. ■

Lemma 17 *We have:*

$$\|\tilde{f}_{\mathbf{W}} - \tilde{\mathbf{f}}_{\mathbf{W}}\| \leqslant \lambda^{-1/2} d_r^{-1} \|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2} q_{\mathbf{W}}\|. \tag{33}$$

Proof The difference $\tilde{f} - \tilde{\mathbf{f}}$ is equal to, with $D = \mathrm{Diag}(\boldsymbol{\zeta}_w^{-1}\mathrm{I})$, $\tilde{f}_{\mathbf{W}} - \tilde{\mathbf{f}}_{\mathbf{W}} = (\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}$. Thus, $\|\tilde{f}_{\mathbf{W}} - \tilde{\mathbf{f}}_{\mathbf{W}}\| \leqslant \lambda^{-1/2}\|D^{-1/2}\|_{\rm op} \times \|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2} q_{\mathbf{W}}\|$, which leads to the desired result. ■

Lemma 18 *Assume* $\|\tilde{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \nu/4$, $\lambda \leqslant |\mathbf{W}| d_r^{-2}$ *and* $\|\boldsymbol{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\rm op} \leqslant \frac{\nu^2\kappa}{16|\mathbf{W}|}$. *We have:*

$$\|\tilde{f}_{\mathbf{W}} - \hat{f}_{\mathbf{W}}\| \leqslant \min\left\{ \frac{96|\mathbf{W}|^{3/2}\|\mathbf{f}_{\mathbf{W}} - \tilde{f}_{\mathbf{W}}\|\Omega(\mathbf{f})^2}{\nu^5\kappa d_r^2}, \frac{\nu^2}{8|\mathbf{W}|^{3/2}}, \frac{\nu}{4} \right\}.$$

Proof We consider the ball of radius $\delta \leqslant \min\{\frac{\nu^2}{8|\mathbf{W}|^{3/2}}, \frac{\nu}{4}\}$ around $\tilde{f}_{\mathbf{W}}$, i.e., $B_\delta(\tilde{f}_{\mathbf{W}}) = \{f_{\mathbf{W}} \in \mathcal{F}_{\mathbf{W}},\ \|f_{\mathbf{W}} - \tilde{f}_{\mathbf{W}}\| \leqslant \delta\}$. Since $\delta \leqslant \nu/4$ and $\|\tilde{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \nu/4$, then in the ball $B_\delta(\tilde{f}_{\mathbf{W}})$, we have for all $w \in \mathbf{W}$, $\|f_{\mathrm{D}(w)\cap\mathbf{W}}\| \geqslant \nu/2$. On the ball $B_\delta(\tilde{f}_{\mathbf{W}})$, the function $L_{\mathbf{W}} : f_{\mathbf{W}} \mapsto L(f_{\mathbf{W}})$ is twice differentiable with Hessian $\widehat{\Sigma}_{\mathbf{WW}}$, while the function $H_{\mathbf{W}} : f_{\mathbf{W}} \mapsto \frac{1}{2}\Omega_{\mathbf{W}}(f_{\mathbf{W}})^2$ is also twice differentiable. The function $H_{\mathbf{W}}$ is the square of a sum of differentiable convex terms; a short calculation shows that the Hessian is greater than the sum of the functions times the sums of the Hessians. Keeping only the Hessians corresponding to the (assumed unique) sources of each of the connected components of $\mathbf{W}$, we obtain the lower bound (which still depends on f):

$$\frac{\partial^2 H_{\mathbf{W}}}{\partial f_{\mathbf{W}} \partial f_{\mathbf{W}}}(f_{\mathbf{W}}) \succcurlyeq d_r \Omega_{\mathbf{W}}(f_{\mathbf{W}})\, \mathrm{Diag}\left[\frac{1}{\|f_C\|}(\mathrm{I} - \|f_C\|^{-2} f_C f_C^\top) \right]_{C \in \mathcal{C}(\mathbf{W})},$$

where $\mathcal{C}(\mathbf{W})$ are the connected components of $\mathbf{W}$. We can now use Lemma 20 to find a lower bound on the Hessian of the objective function $L_{\mathbf{W}} + \lambda H_{\mathbf{W}}$ on the ball $B_\delta(\mathbf{f}_{\mathbf{W}})$: with $A = \lambda_{\min}[(\langle f_C, \widehat{\Sigma}_{CD} f_D\rangle)_{C,D\in\mathcal{C}(\mathbf{W})}]$, we obtain the lower bound

$$B = \frac{A}{3}\min\left\{1, \frac{\lambda d_r^2}{|\mathbf{W}|}\right\} = \frac{A\lambda d_r^2}{3|\mathbf{W}|},$$

because $\Omega_{\mathbf{W}}(f_{\mathbf{W}})\|f_C\|^{-1} \geqslant d_r$, $\lambda_{\max}(\widehat{\Sigma}_{\mathbf{WW}}) \leqslant |\mathbf{W}|$, and $\lambda \leqslant |\mathbf{W}| d_r^{-2}$.

We have moreover on the ball $B_\delta(\tilde{f}_{\mathbf{W}})$ (on which $\|f_{\mathbf{W}}\| \leqslant 2\|\mathbf{f}_{\mathbf{W}}\| \leqslant 2|\mathbf{W}|^{1/2}$),

$$\begin{aligned}
A &\geqslant \lambda_{\min}[(\langle f_C, \boldsymbol{\Sigma}_{CD} f_D\rangle)_{C,D\in\mathcal{C}(\mathbf{W})}] - \max_{C\in\mathcal{C}(\mathbf{W})} \|f_C\|^2 \|\boldsymbol{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\rm op} \\
&\geqslant \kappa \min_{C\in\mathcal{C}(\mathbf{W})} \sum_{w\in C} \|\boldsymbol{\Sigma}_{ww}^{1/2} f_w\|^2 - 4|\mathbf{W}| \|\boldsymbol{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\rm op} \\
&\geqslant \kappa \min_{C\in\mathcal{C}(\mathbf{W})} \sum_{w\in C} \|\boldsymbol{\Sigma}_{ww}^{1/2} \mathbf{f}_w\|^2 - 2\kappa|\mathbf{W}|^{1/2}\delta|\mathbf{W}| - 4|\mathbf{W}| \|\boldsymbol{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\rm op} \\
&\geqslant \kappa\nu^2 - \kappa\nu^2/4 - \kappa\nu^2/4 \geqslant \kappa\nu^2/2,
\end{aligned}$$

because we have assumed that that $2\kappa|\mathbf{W}|^{1/2}\delta|\mathbf{W}| \leqslant \nu^2\kappa/4$ and $4|\mathbf{W}|\|\mathbf{\Sigma_{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}} \leqslant \nu^2\kappa/4$.

We can now show that $\hat{f}_{\mathbf{W}}$ and $\tilde{f}_{\mathbf{W}}$ are close, which is a simple consequence of the lower bound B on the Hessian. Indeed, the gradient of the objective at $\tilde{f}_W$ (applied to z) is equal to

$$\begin{aligned}\langle \nabla_{f_{\mathbf{W}}} L_{\mathbf{W}}(\tilde{f}_{\mathbf{W}}) + \lambda \nabla_{f_{\mathbf{W}}} H_{\mathbf{W}}(\tilde{f}_{\mathbf{W}}), z\rangle &= +\lambda \sum_{v\in\mathbf{W}} \langle (\boldsymbol{\zeta}_w^{-1} - \zeta_w(\tilde{f}_w)^{-1})\tilde{f}_w, z_w\rangle \\ &\leqslant 2\lambda\|z\|\,|\mathbf{W}|^{1/2} \max_{v\in\mathbf{W}} |\boldsymbol{\zeta}_w^{-1} - \zeta_w(\tilde{f}_w)^{-1}| \\ &\leqslant \lambda\|z\|\,|\mathbf{W}|^{1/2} \frac{8\|\mathbf{f_W} - \tilde{f}_{\mathbf{W}}\|}{\nu^3}\Omega(\mathbf{f})^2,\end{aligned}$$

because $|\boldsymbol{\zeta}_w^{-1} - \zeta_w(\tilde{f}_w)^{-1}| \leqslant \frac{2\|\mathbf{f_W} - \tilde{f}_{\mathbf{W}}\|}{\nu\zeta_w} \leqslant \|\mathbf{f_W} - \tilde{f}_{\mathbf{W}}\|\frac{4\Omega(\mathbf{f})^2}{\nu^3}$. If we choose

$$\delta \geqslant 2\frac{\lambda|\mathbf{W}|^{1/2}\frac{8\|\mathbf{f_W}-\tilde{f}_{\mathbf{W}}\|}{\nu^3}\Omega(\mathbf{f})^2}{\frac{\kappa\nu^2}{2}\frac{\lambda d_r^2}{3|\mathbf{W}|}} = \frac{96|\mathbf{W}|^{3/2}\|\mathbf{f_W} - \tilde{f}_{\mathbf{W}}\|\Omega(\mathbf{f})^2}{\nu^5\kappa d_r^2},$$

then the minimum of the reduced cost function must occur within the ball $B_\delta(\tilde{f}_{\mathbf{W}})$. ■

We can now combine the four previous lemma into the following proposition:

Proposition 19 *We have:*

$$\|\tilde{f}_{\mathbf{W}} - \mathbf{f_W}\| \leqslant \left(\lambda + \frac{\|\widehat{\Sigma}_{\mathbf{WW}} - \mathbf{\Sigma_{WW}}\|_{\mathrm{op}}}{d_r^2}\right)\frac{\Omega(\mathbf{f})^2|\mathbf{W}|^{1/2}}{\kappa\nu} + \frac{\lambda^{-1/2}}{d_r}\|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2} q_{\mathbf{W}}\|. \quad (34)$$

Assume moreover $\|\tilde{f}_{\mathbf{W}} - \mathbf{f_W}\| \leqslant \nu/4$, $\lambda \leqslant |\mathbf{W}|d_r^{-2}$ *and* $\|\mathbf{\Sigma_{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}} \leqslant \frac{\nu^2\kappa}{16|\mathbf{W}|}$*; then:*

$$\|\mathbf{f_W} - \hat{f}_{\mathbf{W}}\| \leqslant \|\tilde{f}_{\mathbf{W}} - \mathbf{f_W}\| + \min\left\{\frac{96|\mathbf{W}|^{3/2}\|\tilde{f}_{\mathbf{W}} - \mathbf{f_W}\|\Omega(\mathbf{f})^2}{\nu^5\kappa d_r^2}, \frac{\nu^2}{8|\mathbf{W}|^{3/2}}, \frac{\nu}{4}\right\}. \quad (35)$$

B.7 Global Optimality of the Reduced Solution

We now prove, that the padded solution of the reduced problem $\hat{f}$ is indeed optimal for the full problem if we have the following inequalities (with $\mu = \lambda\Omega(\mathbf{f})d_r$ and $\omega = \Omega(\mathbf{f})d_r^{-1}$):

$$\|\mathbf{\Sigma_{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\| \leqslant \frac{d_r\eta\kappa\nu^2}{10\Omega(\mathbf{f})|\mathbf{W}|^{1/2}} = O\left(\omega^{-1}|\mathbf{W}|^{-1/2}\right) \quad (36)$$

$$\|\mathbf{\Sigma_{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\| \leqslant \frac{\lambda^{1/2}d_r^2\eta\kappa\nu^2}{10\Omega(\mathbf{f})|\mathbf{W}|^{1/2}} = \mu^{1/2}O\left(\omega^{-3/2}|\mathbf{W}|^{-1/2}\right) \quad (37)$$

$$\|\mathbf{f_W} - \hat{f}_{\mathbf{W}}\| \leqslant \frac{\lambda^{-1/2}d_r^2\eta\kappa\nu^5}{40\Omega(\mathbf{f})^3|\mathbf{W}|^{1/2}} = \mu^{-1/2}O\left(\omega^{-5/2}|\mathbf{W}|^{-1/2}\right) \quad (38)$$

$$\|\mathbf{f_W} - \hat{f}_{\mathbf{W}}\| \leqslant \min\left\{\nu\eta/5, \frac{d_r\eta\nu^3}{20\Omega(\mathbf{f})}\right\} = O\left(\omega^{-1}\right) \quad (39)$$

$$\lambda^{1/2} \leqslant \frac{d_r\eta\kappa^{3/2}\nu^3}{20\Omega(\mathbf{f})^2|\mathbf{W}|^{1/2}} \text{ i.e., } \mu^{1/2} = O\left(\omega^{-3/2}|\mathbf{W}|^{-1/2}\right) \quad (40)$$

$$\Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}] \leqslant \lambda\Omega(\mathbf{f})\eta/5 = O(\mu d_r^{-1}) \tag{41}$$

$$\|\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2} q_{\mathbf{W}}\| \leqslant \frac{\lambda d_r^3 \nu^3 \eta}{20\Omega(\mathbf{f})} = \mu O\left(\omega^{-2}\right). \tag{42}$$

Following Appendix A.6, since $\|\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}\| \leqslant \nu/2$, the hull is indeed selected, and $\hat{f}_{\mathbf{W}}$ satisfies the local optimality condition

$$\begin{aligned}
\widehat{\Sigma}_{\mathbf{WW}}(\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}) - q_{\mathbf{W}} + \lambda\Omega_{\mathbf{W}}(\hat{f}_{\mathbf{W}})\hat{s}_{\mathbf{W}} &= 0, \\
\widehat{\Sigma}_{\mathbf{WW}}(\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}}) - q_{\mathbf{W}} + \lambda \operatorname{Diag}(\hat{\zeta}_w^{-1})\hat{f}_{\mathbf{W}} &= 0,
\end{aligned}$$

where $\hat{s}_{\mathbf{W}}$ is defined as (following the definition of $\mathbf{s}$) and $\hat{\zeta} = \zeta(\hat{f}_{\mathbf{W}})$:

$$\hat{s}_w = \Big(\sum_{v \in \mathrm{A}(w)} d_v \|\hat{f}_{\mathrm{D}(v)}\|^{-1} \Big) \hat{f}_w = \hat{\zeta}_w^{-1} \Omega(\hat{f})^{-1} \hat{f}_w, \ \forall w \in \mathbf{W}.$$

This allows us to give a "closed form" solution (not really closed form because it depends on $\hat{\zeta}$, which itself depends on $\hat{f}$):

$$\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}} \quad = \quad (\widehat{\Sigma}_{\mathbf{WW}} + \lambda \operatorname{Diag}(\hat{\zeta}_w^{-1}))^{-1} (q_{\mathbf{W}} - \lambda \operatorname{Diag}(\hat{\zeta}_w^{-1}) \mathbf{f}_{\mathbf{W}}).$$

We essentially replace $\hat{\zeta}$ by $\boldsymbol{\zeta}$ and check the optimality conditions from Appendix A.6. That is, we consider the event $\Omega^*_{\mathbf{W}^c}[\nabla L(\hat{f})_{\mathbf{W}^c}] \leqslant \lambda\Omega(\hat{f})$. We use the following inequality, with the notations $\mathbf{g}_{\mathbf{W}^c} = \operatorname{Diag}(\boldsymbol{\Sigma}_{vv})_{\mathbf{W}^c} C_{\mathbf{W}^c\mathbf{W}} C^{-1}_{\mathbf{WW}} \operatorname{Diag}(\boldsymbol{\Sigma}^{1/2}_{ww}\Omega(\mathbf{f})^{-1}\boldsymbol{\zeta}_w^{-1})_{\mathbf{W}} \mathbf{h}_{\mathbf{W}}$ and $\hat{D} = \operatorname{Diag}(\hat{\zeta}_w^{-1})_{\mathbf{W}}$, $D = \operatorname{Diag}(\boldsymbol{\zeta}_w^{-1})_{\mathbf{W}}$:

$$\begin{aligned}
\Omega^*_{\mathbf{W}^c}[\nabla L(\hat{f})_{\mathbf{W}^c}] &= \Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\hat{f}_{\mathbf{W}} - \mathbf{f}_{\mathbf{W}})] \\
&= \Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda\hat{D})^{-1}(q_{\mathbf{W}} - \lambda\hat{D}\mathbf{f}_{\mathbf{W}})] \\
&\leqslant \Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}] + \lambda\Omega^*_{\mathbf{W}^c}[\mathbf{g}_{\mathbf{W}^c}] \\
&\quad + \lambda\Omega^*_{\mathbf{W}^c}[\mathbf{g}_{\mathbf{W}^c} - \boldsymbol{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}})] \\
&\quad + \lambda\Omega^*_{\mathbf{W}^c}[\boldsymbol{\Sigma}_{w\mathbf{W}}(\boldsymbol{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} D\mathbf{f}_{\mathbf{W}} - \widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda\hat{D})^{-1}\hat{D}\mathbf{f}_{\mathbf{W}}] \\
&\quad + \Omega^*_{\mathbf{W}^c}[\widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda\hat{D})^{-1} q_{\mathbf{W}} - (\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}] \\
&\leqslant \Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}] + \lambda\Omega^*_{\mathbf{W}^c}[\mathbf{g}_{\mathbf{W}^c}] \\
&\quad + \lambda(A + B + C).
\end{aligned}$$

We will bound the last three terms A, B and C by $\Omega(\mathbf{f})\eta/5$, bound the difference $|\Omega(\mathbf{f}) - \Omega(\hat{f})| \leqslant \eta\Omega(\mathbf{f})/5$ (which is implied by Eq. (39)) and use the assumption $\Omega^*_{\mathbf{W}^c}[\mathbf{g}_{\mathbf{W}^c}] \leqslant 1 - \eta$, and use the bound in Eq. (41) to bound $\Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1} q_{\mathbf{W}}] \leqslant \lambda\Omega(\mathbf{f})\eta/5$. Note that we have the bound $\Omega^*_{\mathbf{W}^c}[g_{\mathbf{W}^c}] \leqslant \max_{v \in \mathbf{W}^c} \frac{\|g_v\|}{d_v}$, obtained by lower bounding $\|f_{\mathrm{D}(v)}\|$ by $\|f_v\|$ in the definition of $\Omega_{\mathbf{W}^c}$.

Bounding B. We have:

$$\begin{aligned}
R &= \widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D})^{-1}\hat{D}\mathbf{f}_{\mathbf{W}}) - \mathbf{\Sigma}_{w\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\mathbf{f}_{\mathbf{W}}) \\
&= (\mathbf{\Sigma}_{w\mathbf{W}}-\widehat{\Sigma}_{w\mathbf{W}})(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\mathbf{f}_{\mathbf{W}}) \\
&\quad +\widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D})^{-1}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D-\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D}))((\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\mathbf{f}_{\mathbf{W}})) \\
&\quad +\widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D})^{-1}\operatorname{Diag}(\hat{\zeta}_w^{-1}-\boldsymbol{\zeta}_w^{-1})\mathbf{f}_{\mathbf{W}} \\
\|R\| &\leqslant \|\mathbf{\Sigma}_{w\mathbf{W}}-\widehat{\Sigma}_{w\mathbf{W}}\|_{\mathrm{op}}\|D\|_{\mathrm{op}}|\mathbf{W}|^{1/2}\kappa^{-1} \\
&\quad +\lambda^{-1/2}\|\hat{D}^{-1/2}\|_{\mathrm{op}}\|D\|_{\mathrm{op}}|\mathbf{W}|^{1/2}\kappa^{-1}\left(\|\mathbf{\Sigma}_{\mathbf{WW}}-\widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}}+\lambda\|D-\hat{D}\|_{\mathrm{op}}\right) \\
&\quad +\|D-\hat{D}\|_{\mathrm{op}}|\mathbf{W}|^{1/2} \\
&\leqslant \|\mathbf{\Sigma}-\widehat{\Sigma}\|2\Omega(\mathbf{f})^2\nu^{-2}|\mathbf{W}|^{1/2}\kappa^{-1} \\
&\quad +\lambda^{-1/2}d_r^{-1}\|2\Omega(\mathbf{f})^2\nu^{-2}|\mathbf{W}|^{1/2}\kappa^{-1}\left(\|\mathbf{\Sigma}_{\mathbf{WW}}-\widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{op}}+\lambda 4\Omega(\mathbf{f})^2\nu^{-3}\|\hat{f}-\mathbf{f}\|\right) \\
&\quad +|\mathbf{W}|^{1/2}4\Omega(\mathbf{f})^2\nu^{-3}\|\hat{f}-\mathbf{f}\|,
\end{aligned}$$

which leads to an upper bound $B \leqslant d_r^{-1}\|R\|$. The constraints imposed by Eq. (36), Eq. (37), Eq. (38) and Eq. (39) imply that $B \leqslant \Omega(\mathbf{f})\eta/5$.

Bounding A. We consider the term $\mathbf{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\mathbf{f}_{\mathbf{W}})$. Because of the operator range conditions used by Bach (2008a) and Fukumizu et al. (2007), we can write

$$\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})C_{\mathbf{WW}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})\gamma = \mathbf{\Sigma}_{\mathbf{WW}}\gamma = D\operatorname{Diag}(\mathbf{\Sigma}_{vv})\mathbf{h}_{\mathbf{W}},$$

where $\|\gamma\| \leqslant \|D\|\kappa^{-1}\|\mathbf{h}\|$. We thus have

$$\begin{aligned}
\mathbf{\Sigma}_{w\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\mathbf{f}_{\mathbf{W}} &= \mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})_{\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}D\operatorname{Diag}(\mathbf{\Sigma}_{vv})_{\mathbf{W}}\mathbf{h}_{\mathbf{W}} \\
&= \mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})_{\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}\mathbf{\Sigma}_{\mathbf{WW}}\gamma \\
&= \mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})_{\mathbf{W}}\gamma \\
&\quad -\mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})_{\mathbf{W}}(\mathbf{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}\lambda D\gamma.
\end{aligned}$$

We have moreover

$$\mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}C_{\mathbf{WW}}^{-1}D\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})_{\mathbf{W}}\mathbf{h}_{\mathbf{W}} = \mathbf{\Sigma}_{ww}^{1/2}C_{w\mathbf{W}}C_{\mathbf{WW}}^{-1}C_{\mathbf{WW}}\operatorname{Diag}(\mathbf{\Sigma}_{vv}^{1/2})\gamma,$$

which leads to an upper bound for A:

$$A \leqslant \kappa^{-1/2}\lambda^{1/2}\|D\|_{\mathrm{op}}^{1/2}\|\gamma\| \leqslant \kappa^{-3/2}\lambda^{1/2}\|D\|_{\mathrm{op}}^{3/2}|\mathbf{W}|^{1/2} \leqslant 4\kappa^{-3/2}\lambda^{1/2}\Omega(\mathbf{f})^3\nu^{-3}|\mathbf{W}|^{1/2}.$$

The constraint imposed on Eq. (40) implies that $A \leqslant \Omega(\mathbf{f})\eta/5$.

Bounding C. We consider, for $w \in \mathbf{W}^c$:

$$\begin{aligned}
T &= \widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D})^{-1}q_{\mathbf{W}} - \widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}q_{\mathbf{W}} \\
&= \lambda\widehat{\Sigma}_{w\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1}(D-\hat{D})(\widehat{\Sigma}_{\mathbf{WW}}+\lambda\hat{D})^{-1}q_{\mathbf{W}} \\
\lambda^{-1}\|T\| &\leqslant \lambda^{-1}\|D^{-1}\|_{\mathrm{op}}\|D-\hat{D}\|_{\mathrm{op}}\|(\widehat{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1/2}q_{\mathbf{W}}\| \\
&\leqslant 4\lambda^{-1}d_r^{-2}\Omega(\mathbf{f})^2\nu^{-3}\|\hat{f}_{\mathbf{W}}-\mathbf{f}_{\mathbf{W}}\|\|(\widehat{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1/2}q_{\mathbf{W}}\|,
\end{aligned}$$

leading to the bound $C \leqslant d_r^{-1}\lambda^{-1}\|T\|$. The constraint imposed on Eq. (42) implies that $C \leqslant \Omega(\mathbf{f})\eta/5$.

B.8 Probability of Incorrect Hull Selection

We now need to lower bound the probability of all events from Eq. (36), Eq. (37), Eq. (38), Eq. (39), Eq. (40), Eq. (41) and Eq. (42). They can first be summed up as:

$$\begin{aligned}
\|\mathbf{f}_{\mathbf{W}} - \hat{f}_{\mathbf{W}}\| &\leqslant O\left(\mu^{1/4}\omega^{-1}|\mathbf{W}|^{-1/2}\right)\\
\mu &\leqslant O\left(\omega^{-3}|\mathbf{W}|^{-1}\right)\\
\|\mathbf{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{tr}} &\leqslant O\left(\omega^{-3/2}|\mathbf{W}|^{-1/2}mu^{1/2}\right)\\
\Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1}q_{\mathbf{W}}] &\leqslant \lambda\Omega(\mathbf{f})\eta/5 = O(\mu d_r^{-1})\\
\|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2}q_{\mathbf{W}}\| &\leqslant O\left(\mu^{3/4}\omega^{-1}|\mathbf{W}|^{1/2}\right).
\end{aligned}$$

From Proposition 19, in order to have $\|\mathbf{f}_{\mathbf{W}} - \hat{f}_{\mathbf{W}}\| \leqslant O\left(\mu^{1/4}\omega^{-1}|\mathbf{W}|^{-1/2}\right)$, we need to have $\|\mathbf{f}_{\mathbf{W}} - \tilde{f}_{\mathbf{W}}\| \leqslant O\left(\mu^{1/4}\omega^{-3}|\mathbf{W}|^{-2}\right)$, i.e., $\|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2}q_{\mathbf{W}}\| \leqslant O(\mu^{3/4}\omega^{-7/2}|\mathbf{W}|^{-2})$, $\mu = O(\mu^{1/4}\omega^{-4}|\mathbf{W}|^{-5/2})$ and $\|\mathbf{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{tr}} = O(\mu^{1/4}\omega^{-5}|\mathbf{W}|^{-5/2})$.

From Proposition 15, in order to bound $\|(\widehat{\Sigma}_{\mathbf{WW}}+\lambda D)^{-1/2}q_{\mathbf{W}}\|$, we require $\|\mathbf{\Sigma}_{\mathbf{WW}}-\widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{tr}} = O(\mu^{1/2}\omega^{-1/2}|\mathbf{W}|^{-3/2})$. We finally require the following bounds:

$$\begin{aligned}
\mu &\leqslant O\left(\omega^{-11/2}|\mathbf{W}|^{-7/2}\right)\\
\|\mathbf{\Sigma}_{\mathbf{WW}} - \widehat{\Sigma}_{\mathbf{WW}}\|_{\mathrm{tr}} &\leqslant O\left(\mu^{1/2}\omega^{-3/2}|\mathbf{W}|^{-1/2}\right)\\
\Omega^*_{\mathbf{W}^c}[-q_{\mathbf{W}^c} + \widehat{\Sigma}_{\mathbf{W}^c\mathbf{W}}(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1}q_{\mathbf{W}}] &\leqslant O(\mu d_r^{-1})\\
\|(\widehat{\Sigma}_{\mathbf{WW}} + \lambda D)^{-1/2}q_{\mathbf{W}}\| &\leqslant O\left(\mu^{3/4}\omega^{-7/2}|\mathbf{W}|^{-2}\right).
\end{aligned}$$

We can now use Propositions 14 and 15 as well as Eq. (31) to obtain the desired upper bounds on probabilities.

B.9 Lower Bound on Minimal Eigenvalues

We provide a lemma used earlier in Section B.6.

Lemma 20 *Let Q be a symmetric matrix defined by blocks and (u_i) a sequence of unit norm vectors adapted to the blocks defining Q. We have:*

$$\lambda_{\min}\left(Q + \mathrm{Diag}\left[\mu_i(\mathrm{I} - u_iu_i^\top)\right]\right) \geqslant \frac{\lambda_{\min}[(u_i^\top Q_{ij}u_j)_{i,j}]}{3}\min\left\{1, \frac{\min_i \mu_i}{\lambda_{\max}(Q)}\right\}.$$

Proof We consider the orthogonal complements V_i of u_i, we then have

$$\begin{gathered}
[u_1, \ldots, u_p]^\top\left(Q + \mathrm{Diag}\left[\mu_i(\mathrm{I} - u_iu_i^\top)\right]\right)[u_1, \ldots, u_p] = (u_i^\top Q_{ij}u_j)_{i,j}\\
[V_1, \ldots, V_p]^\top\left(Q + \mathrm{Diag}\left[\mu_i(\mathrm{I} - u_iu_i^\top)\right]\right)[V_1, \ldots, V_p] = (V_i^\top Q_{ij}V_j + \delta_{i=j}\mu_i\mathrm{I})_{i,j}\\
[V_1, \ldots, V_p]^\top\left(Q + \mathrm{Diag}\left[\mu_i(\mathrm{I} - u_iu_i^\top)\right]\right)[u_1, \ldots, u_p] = (V_i^\top Q_{ij}u_j)_{i,j}.
\end{gathered}$$

We can now consider Schur complements: the eigenvalue we want to lower-bound is greater than ν if $\nu \leqslant \lambda_{\min}[(u_i^\top Q_{ij}u_j)_{i,j}]$ and

$$(V_i^\top Q_{ij}V_j + \delta_{i=j}\mu_i\mathrm{I})_{i,j} - (V_i^\top Q_{ij}u_j)_{i,j}((u_i^\top Q_{ij}u_j)_{i,j} - \nu\mathrm{I})^{-1}(u_i^\top Q_{ij}V_j)_{i,j} \succcurlyeq \nu\mathrm{I}$$

which is equivalent to

$$(V_i^\top Q_{ij} V_j)_{i,j} + \mathrm{Diag}(\mu_i \mathrm{I}) - (V_i^\top Q_{ij} u_j)_{i,j} (u_i^\top Q_{ij} u_j)_{i,j}^{-1} (u_i^\top Q_{ij} V_j)_{i,j} \\ + (V_i^\top Q_{ij} u_j)_{i,j} \left[(u_i^\top Q_{ij} u_j)_{i,j}^{-1} - ((u_i^\top Q_{ij} u_j)_{i,j} - \nu \mathrm{I})^{-1} \right] (u_i^\top Q_{ij} V_j)_{i,j} \succcurlyeq \nu \mathrm{I}. \quad (43)$$

If we assume that $\nu \leqslant \lambda_{\min}[(u_i^\top Q_{ij} u_j)_{i,j}]/2$, then the second term has spectral norm less than $\frac{2\nu\lambda_{\max}(Q)}{\lambda_{\min}[(u_i^\top Q_{ij} u_j)_{i,j})}$. The result follows. ■

Acknowledgments

I would like to thank Rodolphe Jenatton, Guillaume Obozinski, Jean-Yves Audibert and Sylvain Arlot for fruitful discussions related to this work. This work was supported by a French grant from the Agence Nationale de la Recherche (MGA Project ANR-07-BLAN-0311).

References

F. Bach. Consistency of the group Lasso and multiple kernel learning. *Journal of Machine Learning Research*, 9:1179–1225, 2008a.

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

F. Bach and M. I. Jordan. Predictive low-rank decomposition for kernel methods. In *Proceedings of the Twenty-second International Conference on Machine Learning (ICML)*, 2005.

F. Bach, G. R. G. Lanckriet, and M. I. Jordan. Multiple kernel learning, conic duality, and the SMO algorithm. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2004a.

F. Bach, R. Thibaux, and M. I. Jordan. Computing regularization paths for learning multiple kernels. In *Advances in Neural Information Processing Systems (NIPS)*, 2004b.

R. Baraniuk. Compressive sensing. *IEEE Signal Processing Magazine*, 24(4):118–121, 2007.

A. Berlinet and C. Thomas-Agnan. *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Klüwer Academic Publishers, 2003.

P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 2009. To appear.

C. L. Blake and C. J. Merz. UCI repository of machine learning databases, 1998.

J.F. Bonnans and A. Shapiro. *Perturbation analysis of optimization problems*. Springer, 2000.

J. M. Borwein and A. S. Lewis. *Convex Analysis and Nonlinear Optimization*. Number 3 in CMS Books in Mathematics. Springer-Verlag, 2000.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2003.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

L. Breiman, J. Friedman, R. Olshen, and C. Stone. *Classification and Regression Trees*. Wadsworth and Brooks, 1984.

H. Brezis. *Analyse Fonctionelle*. Masson, 1980.

P. J. Cameron. *Combinatorics: Topics, Techniques, Algorithms*. Cambridge University Press, 1994.

E. Candès and M. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

O. Chapelle and A. Rakotomamonjy. Second order optimization of kernel parameters. In *NIPS Workshop on Kernel Learning*, 2008.

J. B. Conway. *A Course in Functional Analysis*. Springer, 1997.

M. Cuturi and K. Fukumizu. Kernels on structured objects through nested histograms. In *Advances in Neural Information Processing Systems (NIPS)*, 2006.

A. d'Aspremont, El L. Ghaoui, M. I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–48, 2007.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

R. Diestel. *Graph Theory*. Springer, 2005.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32: 407, 2004.

S. Fine and K. Scheinberg. Efficient SVM training using low-rank kernel representations. *Journal of Machine Learning Research*, 2:243–264, 2001.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

J. H. Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5):1189—1232, 2001.

J. H. Friedman. Multivariate adaptive regression splines. *Annals of Statistics*, 19(1):1–67, 1991.

K. Fukumizu, F. Bach, and A. Gretton. Statistical convergence of kernel canonical correlation analysis. *Journal of Machine Learning Research*, 8(8), 2007.

M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, 2008. `http://www.stanford.edu/~boyd/cvx/`.

K. Grauman and T. Darrell. The pyramid match kernel: Efficient learning with sets of features. *Journal of Machine Learning Research*, 8:725–760, 2007.

C. Gu. *Smoothing Spline ANOVA Models*. Springer-Verlag, 2002.

Z. Harchaoui, F. Bach, and E. Moulines. Testing for homogeneity with kernel Fisher discriminant analysis. Technical Report hal-00270806, HAL, 2008.

T. J. Hastie and R. J. Tibshirani. *Generalized Additive Models*. Chapman & Hall, 1990.

J. Huang, T. Zhang, and D. Metaxas. Learning with structured sparsity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

R. Jenatton, J.-Y. Audibert, and F. Bach. Structured variable selection with sparsity-inducing norms. Technical Report 0904.3523v1, Arxiv, 2009.

V. Koltchinskii and M. Yuan. Sparse recovery in large ensembles of kernel machines. In *Proceedings of the Conference on Learning Theory (COLT)*, 2008.

G. R. G. Lanckriet, T. De Bie, N. Cristianini, M. I. Jordan, and W. S. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20:2626–2635, 2004a.

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

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

C. Lemaréchal and C. Sagastizábal. Practical aspects of the Moreau-Yosida regularization: Theoretical preliminaries. *SIAM Journal on Optimization*, 7(2):867–895, 1997.

Y. Lin and H. H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34(5):2272–2297, 2006.

M. S. Lobo, L. Vandenberghe, S. Boyd, and H. Lébret. Applications of second-order cone programming. *Linear Algebra and its Applications*, 284:193–228, 1998.

H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text classification using string kernels. *Journal of Machine Learning Research*, 2:419–444, 2002.

G. Loosli, S. Canu, S. Vishwanathan, A. Smola, and M. Chattopadhyay. Boîte à outils SVM simple et rapide. *Revue d'Intelligence Artificielle*, 19(4-5):741–767, 2005.

K. Lounici, M. Pontil, A. B. Tsybakov, and S. A. van de Geer. Taking advantage of sparsity in multi-task learning. In *Proceedings of the twenty-second Annual Conference on Learning Theory (COLT)*, 2009.

P. Massart. *Concentration Inequalities and Model Selection: Ecole d'été de Probabilités de Saint-Flour 23*. Springer, 2003.

C. A. Micchelli, Y. Xu, and H. Zhang. Universal kernels. *Journal of Machine Learning Research*, 7:2651–2667, 2006.

Y. Nardi and A. Rinaldo. On the asymptotic properties of the group Lasso estimator for linear models. *Electronic Journal of Statistics*, 2:605–633, 2008.

G. Obozinski, B. Taskar, and M.I. Jordan. Joint covariate selection and joint subspace selection for multiple classification problems. *Statistics and Computing*, 2009. To appear.

B. A. Olshausen and D. J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37:3311–3325, 1997.

C. S. Ong, A. J. Smola, and R. C. Williamson. Learning the kernel with hyperkernels. *Journal of Machine Learning Research*, 6:1043–1071, 2005.

J. Platt. Fast training of support vector machines using sequential minimal optimization. In *Advances in Kernel Methods: Support Vector Learning*, 1998.

M. Pontil and C.A. Micchelli. Learning the kernel function via regularization. *Journal of Machine Learning Research*, 6:1099–1125, 2005.

A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet. SimpleMKL. *Journal of Machine Learning Research*, 9:2491–2521, 2008.

C. E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: Sparse additive models. In *Advances in Neural Information Processing Systems (NIPS)*, 2008.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

V. Roth. The generalized Lasso. *IEEE Transactions on Neural Networks*, 15(1):16–28, 2004.

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

B. Schölkopf and A. J. Smola. *Learning with Kernels*. MIT Press, 2002.

B. Schölkopf, K. Tsuda, and J. P. Vert, editors. *Kernel Methods in Computational Biology*. MIT Press, 2004.

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

S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf. Large scale multiple kernel learning. *Journal of Machine Learning Research*, 7:1531–1565, 2006.

N. Srebro and S. Ben-David. Learning bounds for support vector machines with learned kernels. In *Proceedings of the Conference on Learning Theory (COLT)*, 2006.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of Machine Learning Research*, 2:67–93, 2002.

M. Szafranski, Y. Grandvalet, and A. Rakotomamonjy. Composite kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2008.

G. Szegö. *Orthogonal Polynomials (4th edition)*. American Mathematical Society, 1981.

R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of The Royal Statistical Society Series B*, 58(1):267–288, 1996.

M. Varma and B. R. Babu. More generality in efficient multiple kernel learning. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

S. V. N. Vishwanathan, A. J. Smola, and M. Murty. SimpleSVM. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2003.

G. Wahba. *Spline Models for Observational Data*. SIAM, 1990.

M. J. Wainwright. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (Lasso). *IEEE Transactions on Information Theory*, 2009. To appear.

C. K. I. Williams and M. Seeger. The effect of the input density distribution on kernel-based classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2000.

T. T. Wu and K. Lange. Coordinate descent algorithms for Lasso penalized regression. *Annals of Applied Statistics*, 2(1):224–244, 2008.

Y. Ying and C. Campbell. Generalization bounds for learning the kernel. In *Proceedings of the twenty-second Annual Conference on Learning Theory (COLT)*, 2009.

M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of The Royal Statistical Society Series B*, 68(1):49–67, 2006.

M. Yuan and Y. Lin. On the non-negative garrotte estimator. *Journal of The Royal Statistical Society Series B*, 69(2):143–161, 2007.

T. Zhang. Some sharp performance bounds for least squares regression with ℓ_1 regularization. *Annals of Statistics*, 2009a. To appear.

T. Zhang. On the consistency of feature selection using greedy least squares regression. *Journal of Machine Learning Research*, 10:555—568, 2009b.

P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, To appear, 2009.

H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, December 2006.

H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320, 2005.