A framework for learning affine transformations for multimodal sparse reconstruction

A Framework For Learning Affine Transformations For Multimodal Sparse Reconstruction

João F. C. Mota[a], Evaggelia Tsiligianni[b], and Nikos Deligiannis[b]

[a]Institute of Sensors, Signals and Systems, Heriot-Watt University, Edinburgh, UK
[b]Vrije Universiteit Brussel, Brussels, Belgium, & imec, Leuven, Belgium

ABSTRACT

We introduce a novel framework to reconstruct highly undersampled signals from their measurements using a correlated signal as an aid. The correlated signal, called *side information*, need not be close or similar to the signal to reconstruct. Thus, our framework applies to the case in which the signals are multimodal. We use two main ingredients: the theory of ℓ_1-ℓ_1 minimization, which establishes precise reconstruction guarantees of sparse signals using a *similar* signal as an aid, and a set of training data consisting of several examples of pairs of the signal to reconstruct and the side information. We adopt a statistical framework where the training and the test data are drawn from the same joint distribution, which is assumed *unknown*. Our main insight is that a quantity arising in the ℓ_1-ℓ_1 minimization theory to measure the quality of the side information can be written as the 0-1 loss of a classification problem. Therefore, our problem can be solved with classification methods, such as support vector machines. Furthermore, using statistical learning theory, we provide guarantees for our method. Specifically, the expected value of the side information quality decreases with $O(1/\sqrt{T})$, where T is the number of training samples. Simulations with synthetic data validate our approach.

Keywords: Sparsity, ℓ_1-ℓ_1 minimization, support vector machines, statistical learning.

1. INTRODUCTION

The concept of sparsity allows encoding the structure of certain models in a principled way. During the last decades, it has entered the jargon of the signal processing and statistics communities and has been crucial in the development of many engineering technologies, from data compression [1] to medical imaging [2, 3], gene sequencing [4], remote sensing [5], astronomical imaging [6], and camera design [7].

While encoding structure via sparsity reveals a certain type of prior knowledge about the model, there is another type of prior knowledge that is typically easier to acquire but which, paradoxically, has been used less frequently in reconstruction problems: examples of previously reconstructed signals, either from the same modality or from a different one. For example, in the acquisition and reconstruction of a medical image, we often have access to similar images [8–11]. Some imaging devices acquire different modalities, either sequentially or simultaneously; see, for example, [12–14] for medical imaging, and [15] for remote sensing. And, in video processing and compression, previously reconstructed frames can aid the reconstruction of the current frame [16–19].

A signal that can aid the reconstruction of another signal will be referred to as *prior* or *side information*. Although techniques that effectively integrate prior information into sparse reconstruction schemes have been proposed in recent years [8–11, 20–22], some of which backed up by strong theoretical results [20, 23, 24], they are limited to the case in which the prior information is similar to the signal to reconstruct, and fail when both signals are significantly different, yet correlated. The aim of this paper is to introduce a framework that addresses precisely this case.

Further author information: (Send correspondence to J. F. C. Mota)
J. F. C. Mota: E-mail: j.mota@hw.ac.uk
E. Tsiligianni: E-mail: etsiligi@etrovub.be
N. Deligiannis: E-mail: ndeligia@etrovub.be

Problem statement. Let $x^\star \in \mathbb{R}^n$ be s-sparse, i.e., it has at most s nonzero entries. Suppose we observe m linear measurements $b = Ax^\star$, where the matrix $A \in \mathbb{R}^{m\times n}$ has more columns than rows: $m < n$. Suppose also that we have access to another signal $y \in \mathbb{R}^n$, correlated to $x^\star$, but not necessarily close to it; for example, y need not be sparse and, for a given metric $d : \mathbb{R}^n \to \mathbb{R}$, $d(x^\star - y)$ may be arbitrarily large. Our problem is then: *Reconstruct $x^\star$ from its measurements b using the prior information y as an aid; also, characterize the number of measurements m that guarantee perfect reconstruction.*

Our approach. We address this problem by using a statistical framework that incorporates two main ingredients: a set of training data $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$, which consists of pairs of signals drawn from some unknown joint distribution $\mathcal{P}$, and the theory for ℓ_1-ℓ_1 minimization developed in [23,24]. First proposed in [8], ℓ_1-ℓ_1 minimization adds to the objective function of basis pursuit [25] the ℓ_1-norm of the difference between the optimization variable and the prior information, that is, it solves

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \beta\|x - y\|_1 \\ \text{subject to} & Ax = b\,, \end{array} \tag{1}$$

where $\beta > 0$ is a tradeoff parameter. The work in [23, 24] provides conditions under which $x^\star$ is the unique solution of problem (1). One of such conditions requires the number of measurements to be larger than a quantity $\underline{m}$, which is a function of the sparsity of $x^\star$ and of how close the prior information y is to $x^\star$. Naturally, when y is close to $x^\star$ (in a precise sense defined below), $\underline{m}$ is much smaller than the number of measurements required by basis pursuit [i.e., (1) with $\beta = 0$]; in this case, the prior information helps reconstructing $x^\star$. On the other hand, when y is not similar to $x^\star$, an instance of which is when y is not sparse, $\underline{m}$ becomes large and, as a result, the prior information actually hinders reconstruction, in the sense that (1) may require more measurements than basis pursuit.

In this paper, we propose to modify the prior information via a map $g : \mathbb{R}^n \to \mathbb{R}^n$ to be learned from the training data, and solve instead

$$\begin{array}{ll} \underset{x}{\text{minimize}} & \|x\|_1 + \|x - g(y)\|_1 \\ \text{subject to} & Ax = b\,. \end{array} \tag{2}$$

Notice that we set $\beta = 1$ in (2). The reason is that ℓ_1- ℓ_1 theory states that this is the best choice for β and, hence, we will use this value henceforth. Assuming the training data $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$ and the pair $(x^\star, y)$ are drawn from the same distribution $\mathcal{P}$, our approach generalizes well, as our theoretical and experimental results indicate. We point out that our results are preliminary, as we focus on learning a map g that minimizes just one specific feature of the "quality" of the prior information $g(y)$. To be able to reduce the required number of measurements, we also need to consider other features. This will be explained in detail in Sections 3 and 4.

Organization. We start by briefly overviewing related work in Section 2. Next, Section 3 gives the necessary background on the theory of ℓ_1- ℓ_1 minimization. Our approach is described in Section 4, and experimental results are given in Section 5.

2. RELATED WORK

We divide techniques that perform multi-modality signal reconstruction by learning relations from training data into two categories: coupled dictionary learning, and Bayesian inference. We overview some of this work, and then mention other related approaches.

Coupled dictionary learning. The typical scenario for coupled dictionary learning is an image processing task and several examples of pairs of images from two different modalities. Examples of such tasks include classification [26,27], super-resolution [28–31], image fusion [32,33], source separation [34,35], image retrieval [36], and audio-visual analysis [37]. The main idea in coupled dictionary learning is to explore the degrees of freedom in conventional dictionary learning to enforce coupling between different modalities. Although there are several

ways of enforcing this coupling, to the best of our knowledge, none of them is backed up by theory or has performance guarantees. Also, learning coupled dictionaries is a computationally intensive task.

Bayesian inference. Bayesian approaches to inverse problems also require training data. Typically, a probabilistic model for the signals is postulated, and the model's parameters are learned from training data. Examples of Bayesian approaches involving reconstruction with aid signals include multitask compressive sensing [38, 39], and reconstruction with side information using Gaussian mixture models [40, 41]. These approaches, however, are limited to signals with similar statistical properties. To overcome this limitation, [42, 43] proposed a sparse reconstruction Bayesian approach based on copula functions that leverages correlated signals with distributions potentially different from the distribution of the signal to reconstruct.

Other related work. Related approaches include distributed compressed sensing [44, 45] and simultaneous sparse approximation [46–48]. In these frameworks, multiple correlated signals are reconstructed by enforcing similar sparsity patterns. Note that there is no concept of "learning" or training data in these frameworks.

Finally, we mention that recently there has been considerable interest in approaches that learn inverse operators via deep neural networks, e.g., [49–51], bypassing the need to solve basis pursuit, i.e., (1) with $\beta = 0$, a computationally expensive task. These use several pairs of signals $(x^{(t)}, b^{(t)})$, related by $b^{(t)} = Ax^{(t)}$, as training data. While deep neural networks can also potentially be used in our framework to learn the map g in (2), they would be used in a very different way; we would still need to solve the optimization problem in (2) to perform reconstruction. However, in contrast with [49–51], our approach would not be tied to a specific measurement matrix A. Indeed, the learning aspect of our method refers to the relation between the signal to reconstruct and the side information, and it does not involve any signal measurements.

3. BACKGROUND: ℓ_1-ℓ_1 MINIMIZATION

This section briefly reviews the theory of ℓ_1-ℓ_1 minimization from [23, 24]. There, as in this paper, the goal is to reconstruct an s-sparse signal $x^\star$ from linear measurements $b = Ax^\star$, where $A : \mathbb{R}^{m\times n}$ has a large nontrivial nullspace ($m \ll n$), with the aid of a side information vector $y \in \mathbb{R}^n$. The reconstruction is performed by solving the ℓ_1-ℓ_1 minimization problem in (1).

The work in [23, 24] provides a bound on the number of measurements that guarantees that $x^\star$ can be reconstructed perfectly. The bound should obviously be a function of the quality of the side information, of how close y is to $x^\star$. This notion of quality is captured by two parameters:

$$\ell(x^\star, y) := \big|\{i \,:\, x_i^\star > 0,\ x_i^\star > y_i\} \cup \{i \,:\, x_i^\star < 0,\ x_i^\star < y_i\}\big| \tag{3}$$

$$\xi(x^\star, y) := \big|\{i \,:\, y_i \neq x_i^\star = 0\}\big| - \big|\{i \,:\, y_i = x_i^\star \neq 0\}\big|\,. \tag{4}$$

The parameter ℓ in (3) plays the most prominent role, and was called *number of bad components* in [23, 24]. Notice that it counts the number of components in which $x_i^\star$ is nonzero, and $x_i^\star > y_i$ (if $x_i^\star > 0$) or $x_i^\star < y_i$ (if $x_i^\star < 0$). Therefore, by definition, $\ell(x^\star, y) \leq s$, the sparsity of $x^\star$.

The following result from [24, Thm. 1] considers the case $\beta = 1$, which was shown to be optimal in the sense that it minimizes the bounds for generic β, irrespective of any problem parameter.

Theorem 3.1 ([24])

Let $x^\star \in \mathbb{R}^n$ be the vector to reconstruct and let $y \in \mathbb{R}^n$ be the prior information. Assume $\ell(x^\star, y) > 0$ and that there exists at least one index i for which $x_i^\star = y_i = 0$. Let the entries of $A \in \mathbb{R}^{m\times n}$ be i.i.d. Gaussian with zero mean and variance $1/m$. If

$$m \geq \underline{m} := 2\ell(x^\star, y) \log\left(\frac{n}{s + \xi(x^\star, y)/2}\right) + \frac{7}{5}\left(s + \frac{\xi(x^\star, y)}{2}\right) + 1\,, \tag{5}$$

then, with probability greater than $1 - \exp\big(-\frac{1}{2}(m - \underline{m})^2\big)$, $x^\star$ is the unique solution of (1) with $\beta = 1$.

This result states that if the number of measurements is larger than the quantity in (5), then ℓ_1-ℓ_1 minimization (1) with $\beta = 1$ reconstructs $x^\star$ with high probability, where the probability is over the realizations of the random matrix A. A similar bound had been derived in [52] for basis pursuit, i.e., (1) with $\beta = 0$ or, in other words, without using any side information. Specifically, basis pursuit is guaranteed to reconstruct $x^\star$ under similar assumptions as Theorem 3.1 if $m \geq 2s\log(n/s) + (7/5)s + 1$. Thus, when n is large enough, the bound in (5) can be much smaller the one for basis pursuit. Recall that, by definition, $\ell(x^\star, y) \leq s$.

As shown both in theory and in practice [19, 23, 24], the most important factor in reconstruction using ℓ_1-ℓ_1 minimization is the number of bad components $\ell(x^\star, y)$. Our goal in this paper is therefore to learn a transformation of y, $g(y)$, that renders $\ell\big(x^\star, g(y)\big)$ small. That is the content of the next section.

4. LEARNING TRANSFORMATIONS FROM DATA

This section addresses the problem of learning a map $g : \mathbb{R}^n \to \mathbb{R}^n$ that transforms a possibly unfavorable prior information y into a favorable one, $g(y)$. We start by formulating the problem we want to solve and then show that it is equivalent to solving a set of classification problems. Then, we propose a method for learning g from training data and provide generalization guarantees.

We mention that g will be learned with the goal of minimizing the quantity $\ell(x^\star, y)$ in (3), not the full bound (5). Nevertheless, the framework we introduce here can be extended to minimize the full bound, and this will be the object of future work.

4.1 Problem formulation and relation to classification

The most important term in the bound for ℓ_1-ℓ_1 minimization in (5) is $\ell(x^\star, y)$, so we will focus on learning $g : \mathbb{R}^n \to \mathbb{R}^n$ such that that term is minimized for the training data. More formally, given a training set $\big\{\big(x^{(t)}, y^{(t)}\big)\big\}_{t=1}^T$, our goal is to find a map g that solves

$$\underset{g\in\mathcal{G}}{\text{minimize}} \quad \sum_{t=1}^{T} \ell\Big(x^{(t)}, g\big(y^{(t)}\big)\Big), \tag{6}$$

where $\mathcal{G}$ is a subset of all the maps from $\mathbb{R}^n$ to $\mathbb{R}^n$. This subset encodes our prior knowledge of the problem and should be chosen carefully. Before specifying which family of functions $\mathcal{G}$ we allow, we decompose the problem in (6) into n independent problems.

Decomposing the problem. Representing the ith component of the map $g : \mathbb{R}^n \to \mathbb{R}^n$ as g_i, we first note that $\ell\big(x, g(y)\big)$ can be written as $\ell\big(x, g(y)\big) = \sum_{i=1}^n 1_{\{\text{sign}(x_i)\cdot(x_i - g_i(y))>0\}}$, where 1_E is the indicator function of the event E, i.e., $1_E = 1$ if E holds, and $1_E = 0$ otherwise; cf. the definition of ℓ in (3). Assuming that $\mathcal{G} = \mathcal{G}_1 \times \cdots \times \mathcal{G}_n$, this means that (6) can be written as

$$\underset{g\in\mathcal{G}_1\times\cdots\times\mathcal{G}_n}{\text{minimize}} \quad \sum_{t=1}^{T}\sum_{i=1}^{n} 1_{\left\{\text{sign}\left(x_i^{(t)}\right)\cdot\left(x_i^{(t)} - g_i\left(y^{(t)}\right)\right)>0\right\}}. \tag{7}$$

Switching the order of the sums, (7) decomposes into n independent problems, the ith of which is

$$\underset{g_i\in\mathcal{G}_i}{\text{minimize}} \quad \sum_{t\in\mathcal{T}_i} 1_{\left\{\text{sign}\left(x_i^{(t)}\right)\cdot\left(x_i^{(t)} - g_i\left(y^{(t)}\right)\right)>0\right\}}, \tag{8}$$

where $\mathcal{T}_i := \big\{t : x_i^{(t)} \neq 0\big\}$ is the set of indices in t for which $x_i^{(t)}$ is nonzero. We discarded the terms $t \notin \mathcal{T}_i$ in the objective function of (8), because they impose no constraints on g_i, that is, $t \notin \mathcal{T}_i$ implies $1_{\left\{\text{sign}(x_i^{(t)})\cdot(x_i^{(t)} - g_i(y^{(t)}))>0\right\}} = 0$ for all g_i. We will see next that if we restrict the set of allowable functions $\mathcal{G}_i$ to the affine ones, problem (8) is exactly the empirical risk minimization of a classification problem with the 0-1 loss.

Interpretation as a classification problem. In a binary classification problem [53–56], we are given T sets of features and their respective class, $\{(u^{(t)}, v^{(t)})\}_{t=1}^{T}$, where $u^{(t)} \in \mathbb{R}^n$ represents the features of the tth data point, and $v^{(t)} \in \{-1, +1\}$ its class. The goal is to learn a classifier, i.e., a function $h : \mathbb{R}^n \to \{-1, +1\}$, that separates the feature points according to their class: $h\big(u^{(t)}\big) = v^{(t)}$, for all $t = 1, \ldots, T$. Then, given a data point $u \in \mathbb{R}^n$ not belonging to the training set, h should generalize well, i.e., $h(u)$ should be an accurate prediction of the (unknown) class of u.

One way of learning g from the training data is via *empirical risk minimization*:

$$\underset{h \in \mathcal{H}}{\text{minimize}} \;\; \frac{1}{T} \sum_{t=1}^{T} L\Big(h, (u^{(t)}, v^{(t)})\Big), \tag{9}$$

where $\mathcal{H}$ is a set of allowable functions, also called the hypothesis set, and $L : \mathcal{H} \times (\mathbb{R}^n \times \{-1, +1\}^n) \to \mathbb{R}$ is the loss function. The specific choice for $\mathcal{H}$ and L depends on our prior knowledge about the problem. A natural loss function for classification problems is the 0-1 loss, defined as

$$L^{\text{0-1}}\big(h, (u, v)\big) := \begin{cases} 1 & , \text{ if } h(u) \neq v \\ 0 & , \text{ if } h(u) = v. \end{cases} \tag{10}$$

Regarding the hypothesis set $\mathcal{H}$, many choices are possible, but the simplest ones are generally preferred for their ability to generalize better [53–56]. A simple class of functions that works particularly well in binary classification is the set of affine functions. These are represented by an hyperplane $H_{w,r} := \{z \,:\, w^\top z + r = 0\}$, which is parameterized by $(w, r) \in \mathbb{R}^n \times \mathbb{R}$. A point $u \in \mathbb{R}^n$ is classified as belonging to class 1 if $w^\top u + r > 0$ and to class -1 if $w^\top u + r < 0$.* In this case, the hypothesis set can be written as

$$\mathcal{H} = \Big\{ z \mapsto \operatorname{sign}\big(w^\top z + r\big) \,:\, w \in \mathbb{R}^n,\; r \in \mathbb{R} \Big\}, \tag{11}$$

With this choice for $\mathcal{H}$, the 0-1 loss in (10) can be written more succinctly as $1_{\{v \cdot (w^\top u + r) < 0\}}$. We are now in condition to establish the relation between each problem (8) and binary classification.

Lemma 4.1

If $\mathcal{G}_i$ is the set of hyperplanes in $\mathbb{R}^n$, i.e., $\mathcal{G}_i = H_{\overline{w}_i, \overline{r}_i} := \{y \mapsto \overline{w}_i^\top y + \overline{r}_i \,:\, \overline{w}_i \in \mathbb{R}^n,\; \overline{r}_i \in \mathbb{R}\}$, then problem (8) can be written as an empirical risk minimization problem (9) with the 0-1 loss (10) over the following hypothesis set:

$$\mathcal{H}' = \Big\{ z \mapsto \operatorname{sign}\big(w^\top z + r\big) \,:\, e_i^\top w = 1 \text{ for some } i = 1, \ldots, n,\; w \in \mathbb{R}^n,\; r \in \mathbb{R} \Big\}, \tag{12}$$

where $e_i = (0, \ldots, 1, \ldots, 0) \in \mathbb{R}^n$ is the ith canonical vector.

Notice that the hypothesis set $\mathcal{H}'$ is the same as $\mathcal{H}$ in (11), but with the additional constraint that at least one component of w is 1.

Proof. Since problem (9) with 0-1 loss (10) and hypothesis (11) can be written as

$$\underset{w, r}{\text{minimize}} \;\; \sum_{t=1}^{T} 1_{\{v^{(t)} \cdot (w^\top u^{(t)} + r) < 0\}},$$

*We assume that when $w^\top u + r = 0$, the class of u can be selected arbitrarily. Under a probabilistic model for the data, however, this case will typically happen with probability 0. We will therefore ignore this issue in the subsequent development.

we only need to show that $\operatorname{sign}(x_i^{(t)}) \cdot (x_i^{(t)} - g_i(y^{(t)})) > 0$ with $g_i \in \mathcal{G}_i$ can be written as $v^{(t)} \cdot (w^\top u^{(t)} + r) < 0$, where $u^{(t)}$ and $v^{(t)}$ represent known data, and w and r are parameters of g_i. Indeed, $g_i \in \mathcal{G}_i$ means that $g_i(y^{(t)}) = \overline{w}_i^\top y^{(t)} + \overline{r}_i$, and thus

$$\operatorname{sign}(x_i^{(t)}) \cdot (x_i^{(t)} - \overline{w}_i^\top y^{(t)} - \overline{r}_i) > 0 \quad \Longleftrightarrow \quad \operatorname{sign}(x_i^{(t)}) \cdot \left(\begin{bmatrix} \overline{w}_i^\top & 1 \end{bmatrix} \begin{bmatrix} y^{(t)} \\ -x_i^{(t)} \end{bmatrix} + \overline{r}_i \right) < 0 \,.$$

The correspondence is then immediate: $v^{(t)} = \operatorname{sign}(x_i^{(t)})$, $w = \begin{bmatrix} \overline{w}_i^\top & 1 \end{bmatrix}^\top$, $u^{(t)} = \begin{bmatrix} {y^{(t)}}^\top & -x_i^{(t)} \end{bmatrix}^\top$, and $r = \overline{r}_i$. Notice that this representation requires imposing the last component of w to be 1, and thus justifies using $\mathcal{H}'$ rather than $\mathcal{H}$. □

Drawbacks of empirical risk minimization. We just showed that the problem we want to solve, (7), decomposes into n independent problems, each of which is equivalent to empirical risk minimization for a binary classification problem, with the 0-1 loss function. This approach, however, has two major drawbacks. The first is that minimizing a sum of 0-1 loss functions is a nonconvex, NP-hard problem when the optimal objective is not 0 [57]. The other is that empirical risk minimization fails to generalize well. Indeed, the no-free-lunch theorem of [56, Thm. 5.1] shows that even when the empirical risk minimization is 0, the generalization error can be arbitrarily large.

Overcoming these drawbacks requires two modifications: approximating the 0-1 loss function by a convex surrogate, and regularizing the problem by penalizing hyperplanes that, though valid, are too close to points of any of the classes.

4.2 Classification using SVMs and proposed estimators

We overcome the drawbacks pointed out in the last section by using support vector machines (SVMs) [54–56, 58]. There are essentially two versions of SVMs: one that assumes the training data points are linearly separable (hard-SVM), and another that allows non-linearly-separable data (soft-SVM). Note that a hard-SVM assumes that the optimal value of empirical risk minimization with the 0-1 loss is zero which, in our problem, means that the optimal value of (8) is 0. Although this is unrealistic in some scenarios, we still describe what a hard-SVM is, because it makes the introduction to soft-SVMs easier.

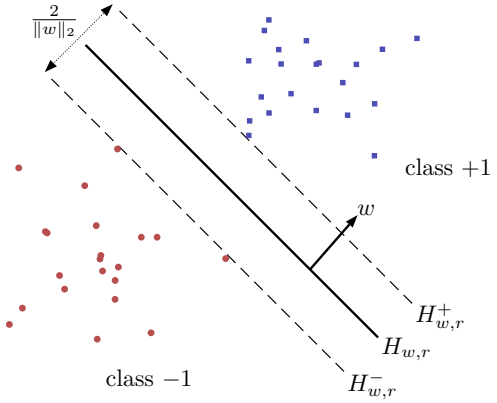


Figure 1. Illustration of an SVM for linearly separable classes: among all the hyperplanes $H_{w,r}$ that separate two different classes of data points, a hard-SVM finds the one that has maximal margin, i.e., it maximizes the distance $2/\|w\|_2$ between $H^-_{w,r}$ and $H^+_{w,r}$.

Hard-SVM. Even when we assume that the optimal value of empirical risk minimization (9) with 0-1 loss (10) and hypothesis (11) is zero, there may exist several solutions (or hyperplanes) that achieve the optimum. Among those, an SVM selects the hyperplane whose margin is maximal. To define margin, consider a hyperplane $H_{w,r} = \{z : w^\top z + r = 0\}$, and two other hyperplanes with the same orientation as $H_{w,r}$, but shifted in opposite

directions: $H^+_{w,r} = \{z : w^\top z + r = 1\}$ and $H^-_{w,r} = \{z : w^\top z + r = -1\}$.[†] See Fig. 1 for an illustration. Then, margin is defined as the distance between $H^+_{w,r}$ and $H^-_{w,r}$, and can be shown to be equal to $2/\|w\|_2$ [[54, §7.1]].[‡] A hard-SVM finds the hyperplane with maximal margin such that all the points in class $+1$ (resp. -1) are supported by $H^+_{w,r}$ (resp. $H^-_{w,r}$), as in Fig. 1. That is, $w^\top u^{(t)} + r \geq 1$ if $v^{(t)} = 1$, and $w^\top u^{(t)} + r \leq -1$ if $v^{(t)} = -1$. In sum, the solution of a hard-SVM is

$$\begin{array}{ll} \underset{w,r}{\text{minimize}} & \|w\|_2^2 \\ \text{subject to} & v^{(t)} \cdot \left(w^\top u^{(t)} + r\right) \geq 1\,, \qquad t = 1, \ldots, T\,, \end{array} \tag{13}$$

a convex problem that can be solved via quadratic programming [59].

Soft-SVM. In practice, the classes of points may not be linearly separable. In this case, not only problem (13) becomes infeasible, but also empirical risk minimization becomes NP-hard [57]. However, a simple modification to (13) can be made to account for this case: introduce a vector $\xi = (\xi_1, \ldots, \xi_T) \in \mathbb{R}^T_+$ of slack variables and allow each constraint of (13) to be violated, while penalizing violations. For example,

$$\begin{array}{ll} \underset{w,r,\xi}{\text{minimize}} & \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{T} 1_T^\top \xi \\ \text{subject to} & v^{(t)} \cdot \left(w^\top u^{(t)} + r\right) \geq 1 - \xi_t\,, \qquad t = 1, \ldots, T \\ & \xi \geq 0_T\,, \end{array} \tag{14}$$

where 0_T, $1_T \in \mathbb{R}^T$ are the all-zeros and all-ones vectors in $\mathbb{R}^T$, and $\lambda > 0$ is a tradeoff parameter. If we view the components of ξ as epigraph variables [59], we can eliminate them and obtain an equivalent formulation of (14):

$$\underset{w,r}{\text{minimize}} \;\; \frac{\lambda}{2}\|w\|_2^2 + \frac{1}{T}\sum_{t=1}^{T} \max\left\{0,\, 1 - v^{(t)} \cdot \left(w^\top u^{(t)} + r\right)\right\}. \tag{15}$$

The function $L^{\text{hinge}}((w,r),(u,v)) := \max\{0, 1 - v \cdot (w^\top u + r)\}$ is known as *hinge loss*. It is convex and, since $\max\{0, 1-x\} \geq 1_{\{x<0\}}$ for all $x \in \mathbb{R}$, it uniformly upper bounds the 0-1 loss: $L^{\text{hinge}}((w,r),(u,v)) = \max\{0, 1 - v \cdot (w^\top u + r)\} \geq 1_{\{v\cdot(w^\top u + r)<0\}} = L^{0\text{-}1}((w,r),(u,v))$, for all $(w,r) \in \mathbb{R}^n \times \mathbb{R}$ and $(u,v) \in \mathbb{R}^n \times \mathbb{R}$. Therefore, (15) is a regularized convex relaxation of (9) with (10)-(11).

Proposed estimators. Given the correspondences established in the proof of Lemma 4.1, we then propose to approximate each problem (8), with $\mathcal{G}_i$ restricted to the set of affine functions, as

$$\underset{\overline{w}_i, \overline{r}_i}{\text{minimize}} \;\; \frac{\lambda}{2}\|\overline{w}_i\|_2^2 + \frac{1}{|\mathcal{T}_i|}\sum_{t \in \mathcal{T}_i} \max\left\{0, 1 - \text{sign}\left(x_i^{(t)}\right) \cdot \left(\overline{w}_i^\top y^{(t)} - x_i^{(t)} + \overline{r}_i\right)\right\}, \tag{16}$$

for $i = 1, \ldots, n$. Notice that problem (16) has to be solved n times, one time for each component. Although (16) is an approximation of our original, nonconvex problem (8), in the next section we show how to bound this approximation error.

4.3 Performance guarantees

In this section, we use a statistical learning framework to provide an upper bound on the expected value of the number of bad components $\ell(x^\star, y)$ in (3). As generalization is impossible without any assumptions on the data, we need to introduce a model of how the test data $(x^\star, y)$ is related to the training data $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1}^T$.

[†]The number 1 in these definitions can be replaced by any other positive constant without any consequence.

[‡]A quick way to see this is to fix $z^- \in H^-_{w,r}$ and compute the margin as the optimal objective of $\min\{\|z - z^-\|_2 : z \in H^+_{w,r}\}$. Squaring the objective and applying the KKT conditions, the solution of this optimization problem is $z^\star = z^- + (1 - r - w^\top z^-)w/\|w\|_2$. Since $-1 - r - w^\top z^- = 0$, we have $\|z^\star - z^-\|_2 = 2/\|w\|_2$.

Assumption 4.2

1. Let $(\Omega, \Sigma, \mathbb{P})$ be a probability space and let $X : \Omega \to \mathbb{R}^n$ and $Y : \Omega \to \mathbb{R}^n$ be random variables whose joint probability distribution is unknown. Assume the training data $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$ and the test datum $(x^\star, y)$ are i.i.d. realizations of the pair (X, Y).
2. There exists $R > 0$ such that $\mathbb{P}\Big\{\|Y\|_2 \leq R\Big\} = 1$.

Our main result is as follows.

Theorem 4.3

Let X and Y denote random vectors in $\mathbb{R}^n$ and let Assumption 4.2 hold. Denote by $W^\star \in \mathbb{R}^{n \times n+1}$ the matrix with rows $\begin{bmatrix} \overline{w}_i^{\star\top} & \overline{r}_i^\star \end{bmatrix}$, $i = 1, \ldots, n$, where $\left(\overline{w}_i^\star, \overline{r}_i^\star\right)$ is a solution of (16). Then, for any $0 < \delta < 1$, with probability at least $1 - \delta$, the expected number of bad components in (3) is bounded as

$$\mathbb{E}\Big[\ell\big(X, Y\big)\Big] \leq \frac{1}{T}\sum_{i=1}^{n}\sum_{t=1}^{T} L^{\text{hinge}}\left(\left(\overline{w}_i^\star, \overline{r}_i^\star\right), \left(y^{(t)}, \text{sign}\Big(x_i^{(t)}\Big)\right)\right) + \frac{2R\|W^\star\|_F^2}{\sqrt{T}} + \left(1 + R\|W^\star\|_F^2\right)\sqrt{\frac{2\log(2/\delta)}{T}}\,. \quad (17)$$

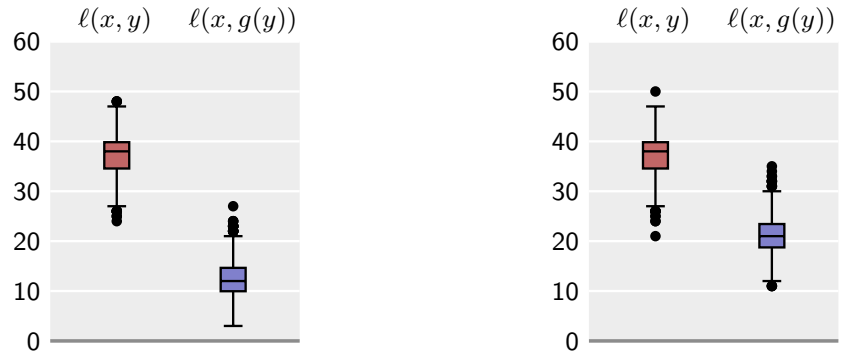
We omit the proof for brevity. It relies on Theorem 26.12 from [56], which is a consequence of McDiarmid's inequality and some properties of the Rademacher complexity measure.

This result is surprisingly general, as it relies on very mild assumptions about the data. It says that, with a given (small) probability over the realizations of the random variables X and Y, the expected value of the number of bad components $\ell(X, Y)$ decreases with $O(1/\sqrt{T})$, where T is the number of training samples. Notice that all the terms in the right-hand side of (17) are known after learning the parameters $\overline{w}_i$ and $\overline{r}_i$ via (16), so the bound can be used in practice to assess the accuracy of estimation. However, as the assumptions are quite general, the bound tends to be loose. An interesting observation is that (17) is independent of λ, but it strongly depends on the norms of $\overline{w}_i$ through the matrix $W^\star$. Indeed, (17) validates the format of the cost function in (16): we attempt to minimize both the norm of $\overline{w}_i$, which affects the last two terms of (17), and the hinge loss of the training data, which contributes to the first term of the right-hand side of (17).

5. EXPERIMENTS

To illustrate the performance of our approach, we conducted simple experiments using synthetic data. Recall that we are focusing on the minimization of the quantity $\ell(x, g(y))$ in (3), not necessarily the global bound (5); hence, our experiments will consider minimizing that quantity only.

Experimental setup. We started by generating the parameters of a true affine transform, $x = W^\star y + r^\star$, with all entries of $W^\star \in \mathbb{R}^{n \times n}$ and $r^\star \in \mathbb{R}^n$ drawn i.i.d. from the normal distribution, and $n = 500$. This rendered an invertible $W^\star$. The training data $\left\{\left(x^{(t)}, y^{(t)}\right)\right\}_{t=1}^{T}$ was generated as follows: each $x^{(t)} \in \mathbb{R}^n$ had a randomly selected support of size 50, and its nonzero entries were drawn i.i.d. normal. The corresponding $y^{(t)}$ was computed as $y^{(t)} = (W^\star)^{-1}(x^{(t)} - r^\star)$. The number of training samples was $T = 2000$. And, using exactly the same procedure, we created an equal number of test samples. The training data was then used to solve the $n = 500$ instances of (16), always with λ set to 0.01. We used CVX [60] as the solver of (16).



(a) Training data (b) Test data

Figure 2. Number of bad components for raw y and for transformed y, $g(y)$, with (a) training data and (b) test data.

Results. The results of the experiments are shown in Fig. 2. Each panel displays two boxplots, the leftmost referring to the values of $\ell(x, y)$, i.e., with raw prior information y, and the rightmost referring to the values of $\ell(x, g(y))$, i.e., with a transformed y. In a boxplot, the median is represented with a horizontal line, and the second and third quartiles are enclosed in a box. The largest (resp. smallest) data point that is at a distance of the upper (resp. lower) quartile smaller than 3/2 times the box height is marked with a "whisker". The points at larger distances are interpreted at outliers and are represented with dots.

Fig. 2(a) shows the boxplots of $\ell(x, y)$ and $\ell(x, g(y))$, with g computed via (16), for the training data. As expected, the values of $\ell(x, g(y))$ were much smaller than the values of $\ell(x, y)$. For example, the median of $\ell(x, y)$ was 38, whereas the median of $\ell(x, g(y))$ was 12.

For the test data, Fig. 2(b) shows similar results, indicating that our approach generalizes well. While the distribution of values of $\ell(x, y)$ is essentially the same as for the training data in Fig. 2(a) [they were generated by a similar process], the values of $\ell(x, g(y))$ are larger for the test data than for the training data. This, of course, is as expected, since the transformation g was learned from the training data. The important point is that in Fig. 2(b) most values of $\ell(x, g(y))$ are smaller that the values of $\ell(x, y)$. Indeed, the median was 38 for the former [the same value as in Fig. 2(a)], and 21 for the latter. This indicates that our approach generalizes well, and has the potential to be used in further research to reduce the global number of measurements in (5).

6. CONCLUSIONS

We introduced a new framework to address the problem of reconstructing sparse signals from a small number of measurements with the aid of side information, a signal correlated with the signal to reconstruct. The main feature of our framework is that it does not require the signal and the side information to be similar and, for this reason, it has the potential to be used in applications with multimodal signals. Our approach uses training data to learn a map that transforms the possibly unfavorable side information into a favorable one. In this paper, we just introduced the framework and focused on improving a specific measure of quality of the side information. The method we propose comes with statistical learning guarantees, and we validated it with simulations on synthetic data.

Our framework opens up several interesting research directions. For example, it can be easily extended to jointly improve the two measures of quality of the side information, as specified in the theory of ℓ_1-ℓ_1 minimization. Another direction is to learn maps more general than just affine ones, for instance, using kernel methods and also neural networks.

REFERENCES

[1] Cover, T. M. and Thomas, J. A., [*Elements of Information Theory*], John Wiley & Sons, Hoboken, New Jersey (1991).

[2] Lustig, M., Donoho, D., and Pauly, J., "Sparse MRI: The application of compressed sensing to rapid MR imaging," *Magnetic Resonance in Medicine* **58**, 1182–1195 (2007).

[3] Davies, M., Puy, G., Vandergheynst, P., and Wiaux, Y., "A compressed sensing framework for magnetic resonance fingerprinting," *SIAM J. Imaging Sciences* **7**(4), 2623–2656 (2014).

[4] Chang, Y. H., Gray, J. W., and Tomlin, C. J., "Exact reconstruction of gene regulatory networks using compressive sensing," *BMC Bioinformatics* **400**(15), 1–22 (2014).

[5] Bioucas-Dias, J. M., Plaza, A., Dobigeon, N., Parente, M., Du, Q., Gader, P., and Chanussot, J., "Hyperspectral unmixing overview: Geometrical, statistical, and sparse regression-based approaches," *IEEE Journal of Selected Topical in Applied Earth Observations and Remote Sensing* **5**(2), 354–379 (2012).

[6] Wiaux, Y., Jacques, L., Puy, G., Scaife, A. M., and Vandergheynst, P., "Compressed sensing imaging techniques for radio interferometry," *Mon. Not. R. Astron. Soc.* **395**, 1733–1742 (2009).

[7] Duarte, M., Davenport, M., Takhar, D., Laska, J., Sun, T., Kelly, K., and Baraniuk, R., "Single-pixel imaging via compressive sampling," *IEEE Sig. Proc. Mag.* **25**(2), 83–91 (2008).

[8] Chen, G.-H., Tang, J., and Leng, S., "Prior image constrained compressed sensing (PICCS): a method to accurately reconstruct dynamic CT images from highly undersampled projection data sets," *Med. Phys.* **35**(2), 660–663 (2008).

[9] Weizman, L., Eldar, Y. C., Eilam, A., Londner, S., Artzi, M., and Ben-Bashat, D., "Fast reference based MRI," in [*Annual Intern. Conf. of the IEEE Engineering in Medicine and Biology Society (EMBC)*], 7586–7489 (2015).

[10] Weizman, L., Eldar, Y. C., and Ben-Bashat, D., "Reference-based MRI," *Med. Phys.* **43**(10), 5357–5369 (2016).

[11] Mota, J. F. C., Weizman, L., Deligiannis, N., Eldar, Y. C., and Rodrigues, M., "Reference-based compressed sensing: A sample complexity approach," in [*IEEE Intern. Conf. Acoustics, Speech, and Sig. Proc. (ICASSP)*], 4687–4691 (2016).

[12] Townsend, D. W., "Multimodality imaging of structure and function," *Phys. Med. Biol.* (53), R1–R39 (2008).

[13] Catana, C., Guimaraes, A. R., and Rosen, B. R., "PET and MR imaging: The odd couple of a match made in heaven," *The Journal of Nuclear Medicine* (54), 815–824 (2013).

[14] Ehrhardt, M. J., Thielemans, K., Pizarro, L., Atkinson, D., Ourselin, S., Hutton, B. F., and Arridge, S. R., "Joint reconstruction of PET-MRI by exploiting structural similarity," *Inverse Problems* **31**(1), 1–23 (2015).

[15] Yuan, X., Tsai, T.-H., Zhu, R., Llull, P., Brady, D., and Carin, L., "Compressive hyperspectral imaging with side information," *IEEE J. Sel. Top. Sig. Proc.* **9**(6), 964–976 (2015).

[16] Wiegand, T., Sullivan, G. J., Bjontegaard, G., and Luthra, A., "Overview of the H.264/AVC video coding standard," *IEEE Transactions on Circuits and Systems for Video Technology* **13**(7), 560–576 (2003).

[17] Vaswani, N., "Kalman filtered compressed sensing," in [*IEEE Intern. Conf. Image Processing (ICIP)*], 893–896 (2008).

[18] Sullivan, G. J., Ohm, J., Han, W. J., and Wiegand, T., "Overview of the high efficiency video coding (HEVC) standard," *IEEE Transactions on Circuits and Systems for Video Technology* **22**(12), 1649–1668 (2012).

[19] Mota, J. F. C., Deligiannis, N., Sankaranarayanan, A. C., Cevher, V., and Rodrigues, M., "Adaptive-rate reconstruction of time-varying signals with application in compressive foreground extraction," *IEEE Trans. Sig. Proc.* **64**(14), 3651–3666 (2016).

[20] Vaswani, N. and Lu, W., "Modified-CS: Modifying compressive sensing for problems with partially known support," *IEEE Trans. Sig. Proc.* **58**(9), 4595–4607 (2010).

[21] Charles, A., Asif, M., Romberg, J., and Rozell, C., "Sparsity penalties in dynamical system estimation," in [*IEEE Conf. Information Sciences and Systems*], 1–6 (2011).

[22] Charles, A. S., Balavoine, A., and Rozell, C. J., "Dynamic filtering of time-varying sparse signals via ℓ_1 minimization," *IEEE Trans. Sig. Proc.* **64**(21), 5644–5656 (2016).

[23] Mota, J., Deligiannis, N., and Rodrigues, M., "Compressed sensing with side information: Geometrical interpretation and performance bounds," in [*IEEE Global Conf. Sig. and Inf. Proc. (GlobalSIP)*], 675–679 (2014).

[24] Mota, J. F. C., Deligiannis, N., and Rodrigues, M. R. D., "Compressed sensing with prior information: Strategies, geometry, and bounds," *IEEE Trans. Inf. Th.* **63**(7), 4472–4496 (2017).

[25] Chen, S., Donoho, D., and Saunders, M., "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comp.* **20**(1), 33–61 (1998).

[26] Wang, K., He, R., Wang, W., Wang, L., and Tan, T., "Learning coupled feature spaces for cross-modal matching," in [*IEEE Conf. Comp. Vision (ICCV)*], 2088–2095 (2013).

[27] Shekhar, S., Patel, V. M., Nguyen, H. V., and Chellappa, R., "Coupled projections for adaptation of dictionaries," *IEEE Trans. Image Processing* **24**(10), 2941–2954 (2015).

[28] Yang, J., Wang, Z., Lin, Z., Cohen, S., and Huang, T., "Coupled dictionary training for image super-resolution," *IEEE Trans. Image Processing* **21**(8), 3467–3478 (2012).

[29] Wang, S., Zhang, L., Liang, Y., and Pan, Q., "Semi-coupled dictionary learning with applications to image super-resolution and photo-sketch synthesis," in [*IEEE Conf. Comp. Vision and Pattern Recognition (CVPR)*], 2216–2223 (2012).

[30] Song, P., Mota, J. F. C., Deligiannis, N., and Rodrigues, M. R. D., "Coupled dictionary learning for multimodal image super-resolution," in [*IEEE Global Conf. Sig. and Inf. Proc. (GlobalSIP)*], 162–166 (2016).

[31] Song, P., Deng, X., Mota, J. F. C., Deligiannis, N., Dragotti, P. L., and Rodrigues, M. R. D., "Joint image super-resolution based on joint sparse representations induced by coupled dictionaries,".

[32] Guo, M., Zhang, H., Li, J., Zhang, L., and Shen, H., "An online coupled dictionary learning approach for remote sensing image fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **7**(4), 1284–1294 (2014).

[33] Song, H., Huang, B., Zhang, K., and Zhang, H., "Spatio-spectral fusion of satellite images based on dictionary-pair learning," *Information Fusion* **18**, 148–160 (2014).

[34] Deligiannis, N., Mota, J. F. C., Cornelis, B., Rodrigues, M. R. D., and Daubechies, I., "X-ray image separation via coupled dictionary learning," in [*IEEE Intern. Conf. Image Processing (ICIP)*], 3533–3537 (2016).

[35] Deligiannis, N., Mota, J. F. C., Cornelis, B., Rodrigues, M. R. D., and Daubechies, I., "Multi-modal dictionary learning for image separation with application in art investigation," *IEEE Trans. Image Processing* **26**(22), 751–764 (2017).

[36] Zhuang, Y., Wang, Y., Wu, F., Zhang, Y., and Lu, W., "Supervised coupled dictionary learning with group structures for multi-modal retrieval," in [*Twenty-Seventh AAAI Conference on Artificial Intelligence*], 1070–1076 (2013).

[37] Monaci, G., Jost, P., Vandergheynst, P., Mailhé, B., Lesage, S., and Gribonval, R., "Learning multimodal dictionaries," *IEEE Trans. Image Processing* **16**(9), 2272–2283 (2007).

[38] Qi, Y., Liu, D., Dunson, D., and Carin, L., "Multi-task compressive sensing with Dirichlet process priors," in [*International Conference on Machine Learning (ICML)*], 768–775 (2008).

[39] Ji, S., Dunson, D., and Carin, L., "Multitask compressive sensing," *IEEE Trans. Sig. Proc.* **57**(1), 92–106 (2009).

[40] Renna, F., Wang, L., Yuan, X., Yang, J., Reeves, G., Calderbank, R., Carin, L., and Rodrigues, M. R. D., "Classification and reconstruction of high-dimensional signals from low-dimensional features in the presence of side information," *IEEE Trans. Inf. Th.* **62**(11), 6459–6492 (2016).

[41] Renna, F., Calderbank, R., Carin, L., and Rodrigues, M. R. D., "Reconstruction of signals drawn from a Gaussian mixture via noisy compressive measurements," *IEEE Trans. Sig. Proc.* **62**(9), 2265–2277 (2014).

[42] Zimos, E., Mota, J. F. C., Rodrigues, M. R. D., and Deligiannis, N., "Bayesian compressed sensing with heterogeneous side information," in [*Data Compression Conference (DCC)*], 191–200 (2016).

[43] Deligiannis, N., Mota, J. F. C., Zimos, E., and Rodrigues, M. R. D., "Heterogeneous networked data recovery from compressive measurements using a copula prior," preprint: http://jmota.eps.hw.ac.uk/ (2017).

[44] Duarte, M., Sarvotham, S., Baron, D., Wakin, M., and Baraniuk, R., "Distributed compressed sensing of jointly sparse signals," *Asilomar Conf. Signals, Systems, and Computers* **24**(11), 1537–1541 (2005).

[45] Baron, D., Wakin, M. B., Duarte, M. F., Sarvotham, S., and Baraniuk, R. G., "Distributed compressed sensing,".

[46] Tropp, J. A., Gilbert, A. C., and Strauss, M. J., "Algorithms for simultaneous sparse approximation. part i: Greedy pursuit," *Signal Processing* **86**(3), 572–588 (2006).

[47] Tropp, J. A., "Algorithms for simultaneous sparse approximation. part i: Convex relaxation," *Signal Processing* **86**(3), 589–602 (2006).

[48] Wipf, D. P. and Rao, B. D., "An empirical Bayesian strategy for solving the simultaneous sparse approximation problem," *IEEE Trans. Sig. Proc.* **55**(7), 3704–3716 (2007).

[49] Gregor, K. and LeCun, Y., "Learning fast approximations of sparse coding," in [*International Conference on Machine Learning (ICML)*], 399–406 (2010).

[50] Mousavi, A. and Baraniuk, R. G., "Learning to invert: Signal recovery via deep convolutional networks,".

[51] Nguyen, D. M., Tsiligianni, E., and Deligiannis, N., "Deep learning sparse ternary projections for compressed sensing of images," submitted (2017).

[52] Chandrasekaran, V., Recht, B., Parrilo, P., and Willsky, A., "The convex geometry of linear inverse problems," *Found. Computational Mathematics* **12**, 805–849 (2012).

[53] Kearns, M. J. and Vazirani, U. V., [*An Introduction to Computational Learning Theory*], The MIT Press, Cambridge, Massachussetts (1994).

[54] Bishop, C., [*Pattern Recognition and Machine Learning*], Springer (2006).

[55] Kulkarni, S. and Harman, G., [*An Elementary Introduction to Statistical Learning Theory*], John Wiley & Sons, Hoboken, New Jersey (2011).

[56] Shalev-Shwartz, S. and Ben-David, S., [*Understanding Machine Learning: From Theory to Algorithms*], Cambridge University Press (2014).

[57] Ben-David, S. and Eiron, N. Long, P. M., "On the dificulty of approximately maximizing agreements," *Journal of Computer and System Sciences* **66**, 496–514 (2003).

[58] Cortes, C. and Vapnik, V., "Support-vector networks," *Machine Learning* **20**, 273–297 (1995).

[59] Boyd, S. and Vandenberghe, L., [*Convex Optimization*], Cambridge University Press (2004).

[60] Grant, M. and Boyd, S., "CVX: Matlab software for disciplined convex programming." http://cvxr.com/cvx (2011).