Structured Regularizers for High-Dimensional Problems: Statistical and Computational Issues

Martin J. Wainwright

Department of Statistics and Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, California 94704; email: wainwrig@stat.berkeley.edu

Annu. Rev. Stat. Appl. 2014. 1:233–53

The *Annual Review of Statistics and Its Application* is online at statistics.annualreviews.org

This article's doi:
10.1146/annurev-statistics-022513-115643

Keywords

regularization, M-estimation, high-dimensional statistics, statistical machine learning, algorithms

Abstract

Regularization is a widely used technique throughout statistics, machine learning, and applied mathematics. Modern applications in science and engineering lead to massive and complex data sets, which motivate the use of more structured types of regularizers. This survey provides an overview of the use of structured regularization in high-dimensional statistics, including regularizers for group-structured and hierarchical sparsity, low-rank matrices, additive and multiplicative matrix decomposition, and high-dimensional nonparametric models. It includes various examples with motivating applications; it also covers key aspects of statistical theory and provides some discussion of efficient algorithms.

1. INTRODUCTION

Regularization has long played a fundamental role in statistics and related mathematical fields. First introduced by Tikhonov (1943) in the context of solving ill-posed integral equations, it has since become a standard part of the statistical tool kit. In the nonparametric setting, various forms of regularization, including kernel smoothing, histogram binning, or penalization, are all used. Moreover, the importance of regularization is only increasing in the modern era of high-dimensional statistics, in which the ambient dimension d of the data may be of the same order or substantially larger than the sample size n. In such high-dimensional settings, for both parametric and nonparametric problems, ill-posedness becomes the rule rather than the exception, so that regularization is essential.

There are as many forms of regularization as there are underlying statistical reasons for using it [for a recent review, see, for instance, Bickel & Li (2006) and accompanying discussion articles]. Given space constraints, this overview is specifically focused on recent advances and open questions involving regularization in the context of high-dimensional M-estimation. The goal is to estimate a quantity of interest (referred to as a parameter) by optimizing the combination of a loss function and a penalty or regularizing function. In this context, regularization serves two purposes, one statistical and the other computational. From a computational perspective, regularization can lend stability to the optimization problem and can lead to algorithmic speed-ups. From a statistical point of view, regularization avoids overfitting and leads to estimators with interesting guarantees on their error. One special case of particular importance is the class of penalized maximum likelihood estimators, but the methodology and theory here described apply more generally. Finally, despite a substantial literature on nonconvex forms of regularization (e.g., see Fan & Li 2001, Zhang & Zhang 2012, and references therein), this overview is limited to convex M-estimators.

This survey begins by introducing the basic idea of a regularized M-estimator. It also includes a brief discussion of some important classical examples, ranging from ridge regression to the Lasso in parametric settings, and including smoothing spline estimates in the nonparametric setting. Section 2 is devoted to a more in-depth discussion of various structured regularizers that are used for estimating vectors, matrices, and functions. Section 3 provides an overview of theory associated with regularized M-estimators in high-dimensional settings. In particular, two properties play an important role in both statistical and algorithmic theory: (*a*) decomposability, a geometric property of the regularizer, and (*b*) restricted strong convexity of the loss function.

1.1. Basic Setup

Let us begin by describing the basic idea of a regularized M-estimator: In brief, it is a method for estimating a quantity of interest θ^* based on solving an optimization problem. In a truly parametric setting, the parameter θ^* corresponds to a finite-dimensional object, such as a vector or a matrix; in a nonparametric setting, it may be an infinite-dimensional object such as a density or regression function. More precisely, let $Z_1^n := \{Z_1, \ldots, Z_n\}$ be a collection of samples drawn with marginal distribution $\mathbb{P}$, and consider an empirical risk function of the form

$$\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(\theta; Z_i), \qquad 1.$$

where the loss function[1] $(\theta; Z_i) \mapsto \mathcal{L}(\theta; Z_i)$ measures the fit of parameter θ to sample Z_i. For instance, in the setting of regression-type data $Z_i = (x_i, y_i)$ with $y_i \in \mathbb{R}$ as the response variable

[1] Note that this use of loss function differs from its classical decision-theoretic use.

and $x_i \in \mathbb{R}^d$ as a covariate vector, a commonly used function is the least-squares criterion $\mathcal{L}(\theta; Z_i) = \frac{1}{2}(y_i - \langle x_i, \theta^* \rangle)^2$. Letting Ω denote the parameter space, the goal is to estimate the parameter $\theta^* \in \Omega$ that uniquely minimizes the population risk $\bar{\mathcal{L}}(\theta) := \mathbb{E}[\mathcal{L}(\theta; Z)]$.

In a regularized M-estimator, the empirical risk function (Equation 1) is combined with a convex regularizer $\mathcal{R} : \Omega \rightarrow \mathbb{R}_+$ that serves to enforce a certain type of structure in the solution. The regularizer and loss function can be combined in one of two ways: A first option is to minimize the loss subject to an explicit constraint involving the regularizer—namely

$$\hat{\theta} \in \arg\min_{\theta \in \Omega} \{\mathcal{L}_n(\theta; Z_1^n)\} \quad \text{subject to } \mathcal{R}(\theta) \leq \rho, \tag{2.}$$

where $\rho > 0$ is a radius to be chosen. A second option is to minimize a weighted combination of the loss and the regularizer

$$\hat{\theta} \in \arg\min_{\theta \in \Omega} \{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta)\}, \tag{3.}$$

where $\lambda_n > 0$ is a regularization weight to be chosen. Under convexity and mild regularity conditions (Bertsekas 1995, Boyd & Vandenberghe 2004), the two families of estimators (Equations 2 and 3) are equivalent, in that for any choice of radius ρ, there is a setting of λ_n for which the solution set of the Lagrangian form (Equation 3) coincides with the constrained form (Equation 2), and vice versa.

1.2. Some Classical Examples

To set the stage for more complicated regularizers, let us begin by considering three classical instances of regularized M-estimators. Perhaps the simplest example of the regularized estimator (Equation 3) is the ridge regression estimate for linear models (Hoerl & Kennard 1970). Given observations of the form $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$ for $i = 1, \ldots, n$, the ridge regression estimator is based on minimizing the weighted sum of the least-squares loss

$$\mathcal{L}_n(\theta; Z_1^n) := \frac{1}{2n} \sum_{i=1}^{n} (y - \langle \theta, x_i \rangle)^2 \tag{4.}$$

coupled with the squared ℓ_2-norm $\mathcal{R}(\theta) = \frac{1}{2}||\theta||_2^2$ as the regularizer. If the responses are generated by a standard linear model of the form $y_i = \langle x_i, \theta^* \rangle + w_i$, where the additive noise w_i is zero-mean with variance σ^2 and independent of the covariates x_i, then the population loss takes the form $\bar{\mathcal{L}}(\theta) = ||\sqrt{\Sigma}(\theta - \theta^*)||_2 + \sigma^2$, where Σ is the covariance matrix of the covariates.

A more recent body of work has focused on the Lasso estimator (Chen et al. 1998, Tibshirani 1996), which replaces the squared ℓ_2 penalty with the ℓ_1-norm regularizer $\mathcal{R}(\theta) = ||\theta||_1 := \sum_{j=1}^{d} |\theta_j|$. Unlike the ℓ_2-regularization of ridge regression, the ℓ_1-penalty promotes sparsity in the underlying solution, which is appropriate when a relatively small subset of covariates are most relevant. There is now an extremely well-developed methodological and theoretical understanding of the Lasso and related ℓ_1-based methods in high-dimensional ($d \gg n$) settings, including its prediction error (e.g., Bickel et al. 2009, Bunea et al. 2007, Greenshtein & Ritov 2004), bounds on the ℓ_2-error (e.g., Bickel et al. 2009, Candes & Tao 2007, Donoho 2006, Donoho & Tanner 2008, Zhang & Huang 2008), and variable selection consistency (e.g., Meinshausen & Bühlmann 2006, Tropp 2006, Wainwright 2009, Zhao & Yu 2006) (for further details and references, see Bühlmann & van de Geer 2011).

Although the linear model is very useful, other prediction problems require richer model classes. In nonparametric regression, the goal is to estimate a function $f : \mathcal{X} \rightarrow \mathbb{R}$ that can be used to predict responses. In many settings, it is natural to seek functions that lie within a Hilbert

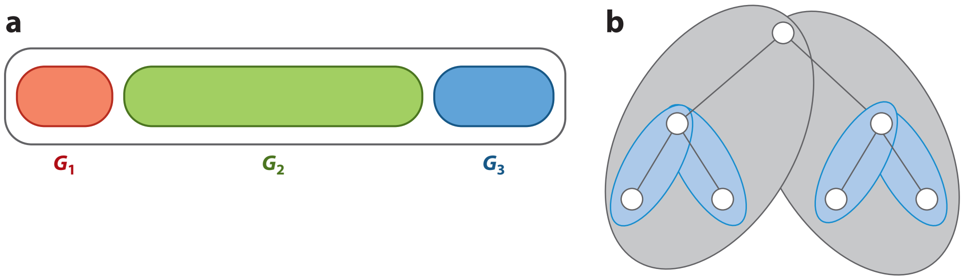

Figure 1

(*a*) Group Lasso penalty with nonoverlapping groups. The groups $\{G_1, G_2, G_3\}$ form a disjoint partition of the index set $\{1, 2, \ldots d\}$. (*b*) A total of $d = 7$ variables are associated with the vertices of a binary tree, and subtrees are used to define a set of overlapping groups. Such overlapping group structures arise naturally in multiscale signal analysis (Bach et al. 2012, Baraniuk et al. 2010).

space $\mathcal{H}$ of functions, meaning a complete inner product space with an associated norm $||f||_{\mathcal{H}}$. Typically, this norm imposes some kind of smoothness condition; for instance, a classical Sobolev smoothness prior for functions $f : [0, 1] \to \mathbb{R}$, given by $||f||_{\mathcal{H}}^2 = f^2(0) + \int_0^1 (f'(t))^2 \mathrm{d}t$, enforces a type of smoothness by penalizing the $L^2[0, 1]$-norm of the first derivative. Given such a norm and observations $Z_i = (x_i, y_i)$ for $i = 1, 2, \ldots, n$, we can then consider estimators of the form

$$\hat{f} \in \arg\min_{f \in \mathcal{H}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda_n \frac{1}{2} ||f||_{\mathcal{H}}^2 \right\}. \qquad 5.$$

There is a wide class of penalized estimators of this type (e.g., Gyorfi et al. 2002, van de Geer 2000), and Section 2.4 discusses certain structured extensions of such smoothing norms that are well suited to high-dimensional problems.

2. STRUCTURED REGULARIZATION AND APPLICATIONS

Let us now turn to an overview of a variety of structured regularizers for different types of high-dimensional problems.

2.1. Group-Structured Penalties

In many applications, a vector or matrix is expected to be sparse, not in an irregular way, but rather in a structured—possibly even hierarchical—manner. To model this type of structured sparsity, researchers have studied various types of group-based regularizers. Consider a collection of groups $\mathcal{G} = \{G_1, \ldots, G_T\}$, where each group is a subset of the index set $\{1, \ldots, d\}$. The union over all groups (see **Figure 1*a*** for one possible grouping of variables) typically covers the full index set, and overlaps among the groups are possible. Given a vector $\theta \in \mathbb{R}^d$, let $\theta_G = \{\theta_s, s \in G\}$ denote the subvector of coefficients indexed by elements of G. Moreover, for each group, let $||\cdot||_G$ denote a norm defined on $\mathbb{R}^{|G|}$. With these ingredients, the associated group norm takes the form

$$\mathcal{R}(\theta) := \sum_{G \in \mathcal{G}} ||\theta_G||_G. \qquad 6.$$

The most common choice is $||\cdot||_G = ||\cdot||_2$ for all groups $G \in \mathcal{G}$, which leads to a norm known as the group Lasso norm (e.g., Kim et al. 2006, Obozinski et al. 2011, Stojnic et al. 2009, Tropp et al. 2006, Yuan & Lin 2006, Zhao et al. 2009). **Figure 2*b*,*c*** provides illustrations of the unit ball

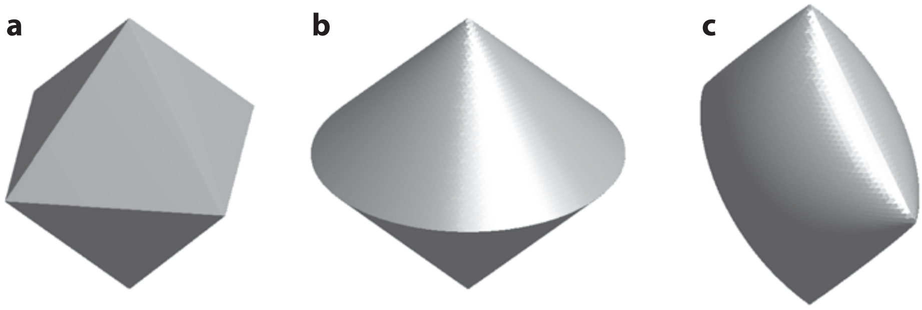

Figure 2

Illustration of unit balls of different norms in $\mathbb{R}^3$. (*a*) The ℓ_1-ball generated by $\mathcal{R}(\theta) = \sum_{j=1}^{3} |\theta_j|$. (*b*) The group Lasso ball generated by $\mathcal{R}(\theta) = \sqrt{\theta_1^2 + \theta_2^2} + |\theta_3|$, corresponding to the groups $G_A = \{1, 2\}$ and $G_B = \{3\}$. (*c*) A group Lasso ball (Equation 6) with overlapping groups, generated by $\mathcal{R}(\theta) = \sqrt{\theta_1^2 + \theta_2^2} + \sqrt{\theta_1^2 + \theta_3^2}$, corresponding to the groups $G_A = \{1, 2\}$ and $G_B = \{1, 3\}$.

of such group Lasso norms in certain cases. The choice $||\cdot||_G = ||\cdot||_\infty$ has also been studied by various authors (e.g., Negahban & Wainwright 2011b, Turlach et al. 2005, Zhao et al. 2009).

In many applications, it is natural to consider structured sparsity with overlapping groups (e.g., Bach et al. 2012, Baraniuk et al. 2010, Jacob et al. 2009, Micchelli et al. 2013, Zhao et al. 2009). For instance, overlapping groups may be defined by the paths in a tree, as shown in **Figure 1*b***. The standard group Lasso norm (Equation 6) is applicable when some of the groups are overlapping—for instance, as with the groups $G_A = \{1, 2\}$ and $G_B = \{1, 3\}$ illustrated in **Figure 2*c***. However, when used in regularized estimators (Equation 2 or 3), the basic group Lasso (Equation 6) has a property that is not always desirable. Given an optimal solution $\hat{\theta}$, let $\hat{S}$ denote its support set—that is, the set of indices for which $\hat{\theta}_j \neq 0$. When using a group-based regularizer, it is often natural to seek optimal solutions $\hat{\theta}$ whose support is given by the union of some subset of the groups. However, if the basic group Lasso (Equation 6) is used as a regularizer, the complement $\hat{S}^c$ of the support—corresponding to elements j for which $\hat{\theta}_j = 0$—is always equal to the union of some subset of the groups. For instance, for the group norm shown in **Figure 2*c***, apart from the full set and empty set, the complement $\hat{S}^c$ can be either $\{1, 2\}$ or $\{1, 3\}$; as a consequence, the support set $\hat{S}$ can be either $\{3\}$ or $\{2\}$, neither of which are unions of subsets of groups.

Motivated to correct this deficiency, Jacob et al. (2009) introduced a variant of the group Lasso. Known as the latent group Lasso, this variant is based on the observation that, for overlapping groups, a vector $\theta \in \mathbb{R}^d$ usually has many possible group representations, meaning collections $\{w_G, G \in \mathcal{G}\}$ such that $\sum_{G \in \mathcal{G}} w_G = \theta$. Minimizing over all such representations yields the following norm:

$$\mathcal{R}(\theta) := \inf_{\substack{\theta = \sum_{G \in \mathcal{G}} w_G \\ w_G, G \in \mathcal{G}}} \left\{ \sum_{G \in \mathcal{G}} ||w_G||_G \right\}, \qquad 7.$$

referred to as the overlapping or latent group Lasso norm. However, when the groups are nonoverlapping, Equation 7 reduces to Equation 6. In the case of overlapping groups, when the overlap group Lasso (Equation 7) is used as a regularizer, any optimal solution $\hat{\theta}$ is guaranteed to have its support $\hat{S}$ equal to a union of groups (Jacob et al. 2009). For instance, returning to the previous example—with groups $G_A = \{1, 2\}$ and $G_B = \{1, 3\}$—the only possible nontrivial supports $\hat{S}$ are $\{1, 2\}$ and $\{1, 3\}$. Thus, its behavior is complementary to the ordinary group Lasso (Equation 6), where these two subsets are only the possible nontrivial complements of the support.

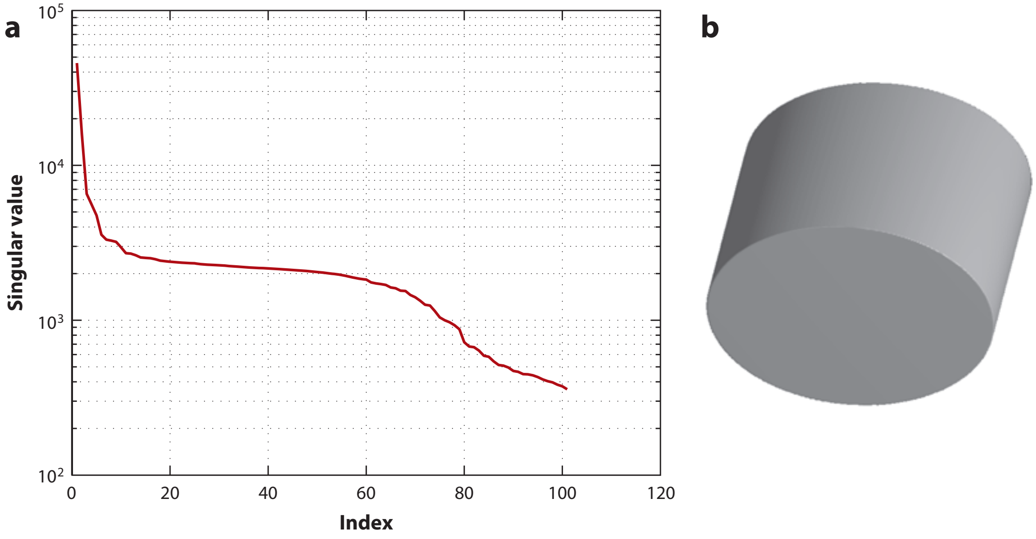

Figure 3

(*a*) Empirical decay of singular values for a subset of the "Jester Joke" database. (*b*) Illustration of the nuclear norm ball as a relaxation of a rank constraint, including a set of all matrices of the form $\Theta = \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}$ such that $|||\Theta|||_{\text{nuc}} \leq 1$. This is a projection of the unit ball of the nuclear norm ball onto the space of symmetric matrices.

2.2. Surrogates to Matrix Rank

A variety of models in multivariate statistics lead to the estimation of matrices with rank constraints. Examples include principal component analysis, canonical correlation analysis, clustering, and matrix completion. In such settings, an ideal approach is often to impose an explicit rank constraint within the estimating procedure. Unfortunately, when viewed as a function on the space of $d_1 \times d_2$ matrices, the rank function is nonconvex, so in many cases, this approach is not computationally feasible.

As one concrete example, the problem of low-rank matrix completion involves noisy observations of a subset of the entries of an unknown matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, and the goal is estimate the full matrix. For instance, this problem arises in collaborative filtering (Srebro et al. 2004, 2005), in which the rows of the matrix correspond to individuals and the columns correspond to items (e.g., books in the Amazon database). The goal is to suggest new items to users on the basis of subsets of items that they and other users have already rated. However, matrix completion problems of this type are ill-posed without some form of structure, and a rank constraint is one natural possibility. For instance, **Figure 3*a*** shows the ordered singular values of a subset of the "Jester Joke" database in which users (rows) rate jokes (columns) in terms of their relative funniness. The rapid drop-off of these singular values confirms empirically that a low-rank approximation is a reasonable model here, as it is in other settings as well (e.g., the Netflix database of movie recommendations). Given a low-rank model, a natural estimator would be to minimize some measure of fit between the data subject to a rank constraint. However, this problem is nonconvex and computationally difficult, so that one is motivated to seek relaxations of it.

The nuclear norm of a matrix provides one natural relaxation of a rank constraint that is analogous to the ℓ_1-norm as a relaxation of the cardinality of a vector. To define the nuclear norm, recall the singular value decomposition (SVD) of a matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$. Letting $d = \min\{d_1, d_2\}$, the SVD takes the form $\Theta = UDV^T$, where $U \in \mathbb{R}^{d_1 \times d}$ and $V \in \mathbb{R}^{d_2 \times d}$ are both orthogonal matrices ($U^T U = V^T V = I_{d \times d}$). The matrix $D \in \mathbb{R}^{d \times d}$ is a diagonal matrix with its entries corresponding to the singular values of Θ, namely the sequence of non-negative numbers

$$\sigma_1(\Theta) \geq \sigma_2(\Theta) \geq \sigma_3(\Theta) \geq \cdots \geq \sigma_d(\Theta) \geq 0. \qquad 8.$$

The number of strictly positive singular values specifies the rank, i.e., $\operatorname{rank}(\Theta) = \sum_{j=1}^{d} \mathbb{I}[\sigma_j(\Theta) > 0]$. This observation, though not practically useful on its own, suggests a natural convex relaxation of a rank constraint, namely the nuclear norm

$$|||\Theta|||_{\text{nuc}} := \sum_{j=1}^{d} \sigma_j(\Theta) \qquad 9.$$

corresponding to the ℓ_1-norm of the singular values (for which no absolute value is necessary because singular values are non-negative by definition). Unlike the matrix rank, the nuclear norm is a convex function on the space of matrices. **Figure 3*b*** provides an illustration of the unit ball of this nuclear norm in a simple case. The statistical and computational behavior of the nuclear norm as a regularizer has been studied for various models, including recovery of matrices from linear projections (e.g., Candès & Plan 2011, Fazel 2002, Negahban & Wainwright 2011a, Recht et al. 2010), more general forms of matrix regression (e.g., Bach 2008, Negahban & Wainwright 2011a, Rohde & Tsybakov 2011), as well as matrix completion (e.g., Candès & Recht 2009, Gross 2011, Koltchinskii et al. 2011, Mazumder et al. 2010, Negahban & Wainwright 2012, Recht 2011, Srebro et al. 2005).

The nuclear norm also has an interesting variational representation in terms of a penalized matrix factorization. In particular, for a given matrix $\Theta \in \mathbb{R}^{d_1 \times d_2}$, suppose that we consider all possible factorizations of the form $\Theta = AB^T$, where $A \in \mathbb{R}^{d_1 \times m}$ and $B \in \mathbb{R}^{d_2 \times m}$ for some arbitrary positive integer m. Suppose that we penalize the components (A, B) of this factorization in terms of their Frobenius norm $|||A|||_{\text{F}} = \sqrt{\sum_{j=1}^{d_1} \sum_{k=1}^{m} A_{jk}^2}$, with the Frobenius norm of B similarly defined. If we minimize the product of the Frobenius norms over all factorizations, it is an elementary exercise to show that we recover the nuclear norm, e.g.,

$$|||\Theta|||_{\text{nuc}} = \inf_{\Theta = AB^T} |||A|||_{\text{F}} |||B|||_{\text{F}}. \qquad 10.$$

This variational representation is useful, in that it leads to alternative algorithms for optimization with the nuclear norm, ones which operate in the space of penalized components.

In addition, the variational representation (Equation 10) suggests other forms of penalized matrix factorization based on replacing the Frobenius norm penalty with other matrix norms. For instance, the matrix max-norm is given by

$$|||\Theta|||_{\max} = \inf_{\Theta = AB^T} |||A|||_{2 \to \infty} |||B|||_{2 \to \infty}, \qquad 11.$$

where $|||A|||_{2 \to \infty} := \max_{j = 1, \ldots, d_1} ||A_j||_2$ is the maximum ℓ_2-norm taken over all rows of A, with $|||B|||_{2 \to \infty}$ defined similarly. The max-norm and other related matrix norms have been studied by Srebro et al. (2004, 2005).

2.3. Structured Norms for Additive Matrix Decomposition

This section describes matrix decompositions that have an additive (as opposed to multiplicative) form. In the most basic form of matrix decomposition, one makes noisy observations of an unknown matrix that has an additive decomposition of the form $\Theta^* = A^* + B^*$. However, this problem is ill-defined unless the matrices A^* and B^* are somehow restricted, and one possibility is that A^* is a low-rank matrix (or is well approximated by one), whereas B^* is sparse in a certain sense. There are various statistical motivations for such "sparse plus low-rank" decompositions, a few of which are considered below.

2.3.1. Robust forms of matrix completion and principal component analysis. Recall the problem of matrix completion, as described in Section 2.2. Suppose that we are performing matrix completion in the context of recommender systems (e.g., Amazon's rating system for books) and that a subset of people wish to selectively alter the output of a recommendation system (e.g., a group of authors who may want their books to be highly recommended). Such individuals may create a fake user account (indexed by a row of the matrix) and then populate that row with selectively chosen rankings.[2] It would then be appropriate to augment our low-rank model for the true ranking matrix (A^* in this case) with a sparse component B^* that had relatively few nonzero rows. Closely related are various forms of robust principal component analysis (PCA) in which an observed data matrix is modeled as consisting of a low-rank part with some form of adversarial but sparse noise (e.g., Candès et al. 2011, Hsu et al. 2011, Xu et al. 2012).

2.3.2. Gaussian graphical models with hidden variables. Another interesting example [studied in depth by Chandrasekaran et al. (2012)] concerns Gaussian graphical model selection with hidden variables. Consider a $(d+r)$-dimensional jointly Gaussian random vector $(X_1, \ldots, X_{d+r})$. The conditional independence properties of such a Gaussian random vector are reflected in the structure of its inverse covariance matrix. In particular, the inverse covariance matrix has a zero in position (s, t) if and only if X_s is conditionally independent of X_t given the collection $\{X_u, u \notin \{s, t\}\}$. When all $(d+r)$-elements of the random vector X are fully observed, there are a variety of statistical estimators designed to exploit this sparsity (e.g., Cai et al. 2011, 2012; Friedman et al. 2008; Lam & Fan 2009; Meinshausen & Bühlmann 2006; Ravikumar et al. 2011; Rothman et al. 2008; Yuan 2010; Zhou et al. 2008). Now suppose that only the first d-components of the random vector are observed, whereas the remaining r components remain hidden or unobserved. In this case, the inverse covariance matrix of $(X_1, \ldots, X_d)$—denoted by $\Theta^* \in \mathbb{R}^{d\times d}$—is no longer sparse in general, because conditional independence properties can be destroyed by the effect of marginalizing out the remaining r hidden variables. In the Gaussian case, using the block matrix inversion formula, the matrix Θ^* can be decomposed additively in terms of a sparse component A^* and a low-rank perturbation B^* guaranteed to have rank at most r, corresponding to the number of hidden variables. This additive decomposition can be exploited to develop effective estimators (Chandrasekaran et al. 2012).

In these and other applications of low-rank plus sparse matrix decomposition, it is useful to consider matrix norms of the form

$$\mathcal{R}(\Theta) := \inf_{\substack{A,B \\ \Theta = A+B}} \{||A||_1 + \omega |||B|||_{\text{nuc}}\} \qquad 12.$$

for a suitably chosen weight $\omega > 0$. These norms can be viewed as a generalized form of overlap group Lasso norm (Equation 7), in which the "groups" correspond to the matrix elements (ℓ_1-norm) and the matrix singular values (nuclear norm). Matrix decomposition in the noiseless setting using the norm (Equation 12), since first being proposed by Chandrasekaran et al. (2011), has

[2] Exactly such a manipulation occurred in 2001, when some adversarial users manipulated the Amazon recommender system so that it would suggest a sex manual to people who enjoyed Christian spiritual guides.

been extensively studied in both the noiseless (e.g., Candes et al. 2011, Hsu et al. 2011, Xu et al. 2012) and noisy settings (e.g., Agarwal et al. 2012b, Hsu et al. 2011).

In addition to the pairing of the ℓ_1- and nuclear norms, other forms of the composite norm (Equation 12) have been studied. For instance, Jalali et al. (2010) studied a combination of the ℓ_1-norm and the blockwise ℓ_1/ℓ_∞-norm. They proved that it has an interesting adaptivity property in terms of variable selection performance: When the variables have block-structured sparsity, the composite norm relaxation achieves the optimal rate (that would be achieved by the ℓ_1/ℓ_∞ penalty alone), but it also achieves the optimal ℓ_1-based rate when there is no block sparsity. In contrast, the ℓ_1/ℓ_∞-norm alone does not have this type of adaptivity (Negahban & Wainwright 2011b) and can perform more poorly than ℓ_1-norm relaxation in the absence of block sparsity.

2.4. Structured Hilbert Norms

Recall the problem of nonparametric regression over Hilbert spaces first discussed in Section 1.1. The unstructured version of this problem—as with all nonparametric problems—suffers severely from the "curse of dimensionality," meaning that they require a sample size that grows exponentially in the dimension. For example, given the space of all twice-differentiable regression functions, obtaining an estimate of accuracy δ in mean-squared error requires a sample size n of the order $(1/\delta)^{1+\frac{d}{4}}$. This fact can be confirmed by inverting known minimax lower bounds on estimation error for twice-differentiable regression functions (Stone 1982).

Accordingly, it is essential to study regression models that have more structure, and these models become particularly interesting in the high-dimensional setting. Stone (1985) introduced a class of additive models in which the regression function $f^* : \mathbb{R}^d \to \mathbb{R}$ is assumed to have an additive decomposition of the form $f^*(x_1, \ldots, x_d) = \sum_{j=1}^d f_j^*(x_j)$, where each f_j belongs to some univariate function space $\mathcal{H}_j$. Here the curse of dimensionality is mostly circumvented: Instead of an exponential growth, the sample size need grow only linearly in dimension. For applications with $n < d$, even more structure is required, and one possible extension is the sparse additive model, in which we assume that only a subset S of the coordinates are associated with nonzero functions. More precisely, consider functions of the form

$$f^*(x_1, \ldots, x_d) = \sum_{j \in S} f_j^*(x_j) \quad \text{with} \quad f_j^* \in \mathcal{H}_j \quad \text{for each} \quad j, \qquad 13.$$

where $S \subset \{1, 2, \ldots, d\}$ is some unknown subset of cardinality s. This class of models, which has been extensively studied over the past decade (e.g., Koltchinskii & Yuan 2008, 2010; Lin & Zhang 2006; Meier et al. 2009; Raskutti et al. 2012; Ravikumar et al. 2009), can be viewed as a semiparametric extension of the sparse linear model, where the unknown functions $\{f_j^*\}_{j=1}^d$ constitute the nonparametric component and the unknown subset S is the parametric component. Liu et al. (2009) described an interesting application to non-Gaussian graphical models involving univariate copulas.

When each $\mathcal{H}_j$ is some univariate Hilbert space, the most natural regularizer associated with the SpAM (sparse additive model) (Equation 13) is based on composing the Hilbert norm with the ℓ_1-norm, which yields the ℓ_1-Hilbert norm $||f||_{\mathcal{H},1} := \sum_{j=1}^d ||f_j||_{\mathcal{H}}$. Given a collection of samples $\{(x_i, y_i)\}_{i=1}^n$, another way in which sparsity can be enforced is a composite ℓ_1-penalty based on the empirical $L^2(\mathbb{P}_n)$-norm, namely the quantity

$$||f||_{n,1} := \sum_{j=1}^d ||f_j||_n, \qquad 14.$$

where $||f_j||_n := \sqrt{\frac{1}{n}\sum_{i=1}^n f_j^2(x_{ij})}$.Both forms of regularization have been studied by different researchers, and it is useful to consider the broader family of regularized M-estimators

$$\hat{f} \in \arg \min_{\substack{f=\sum_{j=1}^d f_j \\ f_j \in \mathcal{H}_j}} \left\{ \frac{1}{2n} \sum_{i=1}^n \Big(y_i - \sum j = 1^d f_j(x_{ij})\Big)^2 + \lambda_n ||f||_{\mathcal{H},1} + \mu_n ||f||_{n,1} \right\}, \qquad 15.$$

parameterized by two non-negative regularization weights λ_n and μ_n, one for each of the sparsity-promoting norms. Several authors (e.g., Koltchinskii & Yuan 2008, Lin & Zhang 2006) have studied the special case of this estimator with $\mu_n = 0$ and a positive choice of λ_n, whereas others (e.g., Meier et al. 2009, Ravikumar et al. 2009) have studied methods closely related to the version with $\lambda_n = 0$ and a positive choice of μ_n. Koltchinskii & Yuan (2010) and Raskutti et al. (2012) studied the fully general class of estimators, with minimax-optimal rates established by the latter.

3. SOME THEORY

We now turn to some theory, both statistical and computational, for regularized M-estimators. Owing to space constraints, discussion is limited to bounds on the statistical estimation error and to certain optimization algorithms. Discussion includes the key ingredients involved in obtaining bounds on the error $\hat{\theta} - \theta^*$ between any optimum of the regularized M-estimator (Equation 3) and the unknown parameter of interest. These bounds are nonasymptotic and thus illustrate the scaling of the required sample size as a function of the problem dimension and other structural quantities. We then discuss some optimization algorithms for solving both the constrained (Equations 2 and 3) problems and consider how the same conditions used to control statistical error can also be used to control optimization error.

3.1. Curvature and Restricted Strong Convexity

Recall that the estimators (Equations 2 and 3) are formulated in terms of optimization. In the classical setting of maximum likelihood, the loss function is given by the negative log-likelihood. The Hessian of the negative log-likelihood evaluated at θ^* corresponds to the Fisher information, which by classical asymptotic theory governs the accuracy of the maximum likelihood estimate. In geometric terms, the Hessian of the loss function encodes the curvature of the loss function, and hence the relative distinguishability of the parameters. In the high-dimensional setting ($d \gg n$), a uniform lower bound on the curvature can no longer be expected. Rather, the generic picture of a high-dimensional convex loss is given in **Figure 4*a***: Although it exhibits positive curvature in certain directions, there is a very large space—at least $d - n$ dimensional—in which it is entirely flat. Thus, if we expect to obtain nontrivial bounds on the statistical error $\hat{\theta} - \theta^*$, the role of regularization is clear: It must exclude certain directions in space, leaving only directions with positive curvature.

How to formalize this geometric intuition analytically? If we require only first derivatives, a natural way to do so is by examining the error in the first-order Taylor series expansion, evaluated at θ^*. In particular, for each direction Δ such that $(\theta^* + \Delta) \in \Omega$, the quantity

$$\mathcal{T}(\Delta; \theta^*, Z_1^n) := \mathcal{L}_n(\theta^* + \Delta; Z_1^n) - \big\{\mathcal{L}_n(\theta^*; Z_1^n) + \langle \nabla \mathcal{L}_n(\theta^*; Z_1^n), \Delta \rangle\big\} \qquad 16.$$

represents the difference between $\mathcal{L}_n(\theta^* + \Delta; Z_1^n)$ and the first-order tangent approximation. This difference is always non-negative for a convex loss function, and for a strongly convex function, it is lower bounded by a quadratic form, uniformly over all Δ.

Instead of such a uniform lower bound, let us consider a relaxed version, formulated in terms of a given norm $\|\cdot\|$, often the Euclidean norm for a vector or the Frobenius norm for a matrix. In

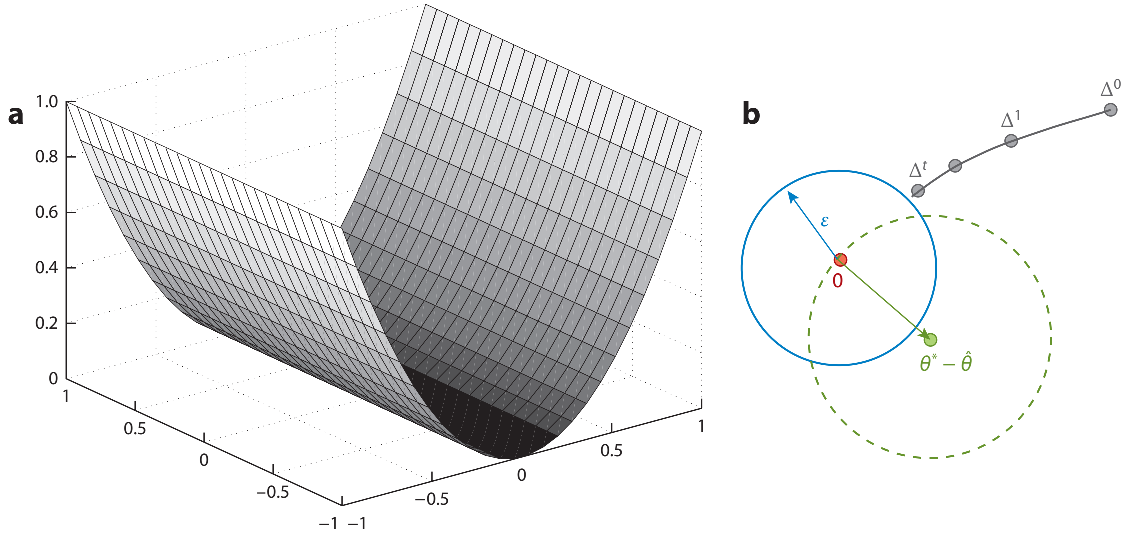

Figure 4

(*a*) Generic illustration of a convex loss function in the high-dimensional ($d \gg n$) setting (Negahban et al. 2012). It is strongly convex in certain directions but completely flat in other directions. The purpose of regularization is to exclude the flat directions. (*b*) Geometry of the error bound (Equation 30). The optimization error $\Delta^t := \theta^t - \hat{\theta}$ decreases rapidly up to a certain tolerance ε. As long as this tolerance is of the same (or lower) order than the statistical precision $||\hat{\theta} - \theta^*||$, the algorithm's output is as good as the global optimum (Agarwal et al. 2012a).

such cases, the loss function $\mathcal{L}_n$ is said to satisfy a restricted strong convexity condition (Negahban et al. 2012), with curvature $\kappa > 0$ and tolerance $\tau_n \geq 0$, with respect to the regularizer $\mathcal{R}$ if

$$\mathcal{T}(\Delta; \theta^*, Z_1^n) \geq \kappa ||\Delta||^2 - \tau_n \mathcal{R}^2(\Delta) \quad \text{for all } ||\Delta|| \leq 1. \tag{17}$$

If such a condition holds with tolerance $\tau_n = 0$, then it is equivalent to a strong convexity condition on the loss. But as discussed above in the high-dimensional setting, strong convexity does not hold. With a strictly positive tolerance $\tau_n > 0$, the restricted strong convexity (RSC) condition (Equation 17) is much milder: It imposes only a nontrivial constraint for vectors Δ such that

$$\frac{\mathcal{R}^2(\Delta)}{||\Delta||^2} \leq \frac{\kappa}{\tau_n}. \tag{18}$$

Equation 17 takes a particularly simple form for the least-squares loss (Equation 4) and samples $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$, for which an elementary calculation shows that the Taylor series error (Equation 16) is given by $\mathcal{T}(\Delta; \theta^*, Z_1^n) = \frac{1}{2n} ||X\Delta||_2^2$. Here $X \in \mathbb{R}^{n \times d}$ denotes the design matrix with $x_i \in \mathbb{R}^d$ as its ith row. Consequently, restricted strong convexity for least-squares loss reduces to a bound of the form

$$\frac{1}{2n} ||X\Delta||_2^2 \geq \kappa ||\Delta||^2 - \tau_n \mathcal{R}^2(\Delta) \quad \text{for all } \Delta \in \mathbb{R}^d. \tag{19}$$

In this way, it is equivalent to lower bounding the restricted ℓ_2-eigenvalues (Bickel et al. 2009) of the sample covariance matrix $\hat{\Sigma}_n = X^T X / n$ (for a thorough discussion of such restricted eigenvalue conditions and their properties, see van de Geer & Buhlmann 2009).

For various classes of random design matrices, forms of the RSC condition (Equation 19) hold with high probability. For instance, for a random design matrix X with rows $x_i \in \mathbb{R}^d$ drawn independently and identically distributed (i.i.d.) from a zero-mean sub-Gaussian distribution with covariance Σ, the following forms of restricted strong convexity hold:

- For $\mathcal{R}(\Delta) = ||\Delta||_1$ and the norm $||\Delta|| := ||\sqrt{\Sigma}\Delta||_2$, Equation 19 holds with high probability with curvature $\kappa = 1/2$ and tolerance $\tau_n \asymp \frac{\log d}{n}$. For results of this type, the reader is referred to Raskutti et al. (2010) and Rudelson & Zhou (2012). This result implies a bound on the ℓ_2-restricted eigenvalues of the matrix X; in particular, from the inequality in Equation 18, the null space cannot contain any vectors $\Delta \in \mathbb{R}^d$ that are relatively sparse in the sense that $\frac{||\Delta||_1^2}{||\sqrt{\Sigma}\Delta||_2^2} \precsim \frac{n}{\log d}$. For further details on restricted eigenvalue and null-space properties, the reader is referred to Bickel et al. (2009), Cohen et al. (2009), Raskutti et al. (2010), and van de Geer & Buhlmann (2009).
- For the group Lasso norm (given by Equation 6) with $|\mathcal{G}|$-groups with maximum group size $g_{\max}$ and the norm $||\Delta|| := ||\sqrt{\Sigma}\Delta||_2$, Equation 19 holds with high probability with curvature $\kappa = 1/2$ and tolerance $\tau_n \asymp \frac{g_{\max}}{n} + \frac{\log|\mathcal{G}|}{n}$. For results of this type, the reader is referred to Negahban et al. (2012).
- Given a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$, the problem of matrix regression involves noisy observations of the form

  $$y_i = \text{trace}(X_i^T \Theta^*) + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{20.}$$

  where $X_i \in \mathbb{R}^{d_1 \times d_2}$ is an observation matrix. (For instance, in the matrix completion problem, the matrix X_i is a "mask," with a single 1 in the position corresponding to the observed entry and zeros elsewhere.) For the nuclear norm $\mathcal{R}(\Delta) = |||\Delta|||_{\text{nuc}}$ and Frobenius norm $||\Delta|| = |||\Delta|||_{\text{F}}$, Equation 19 holds with curvature $\kappa = 1/2$ and tolerance $\tau_n \asymp \frac{d}{n}$. For results of this type, the reader is referred to Candès & Plan (2011) and to Negahban & Wainwright (2011a, 2012).

Forms of restricted strong convexity also hold for more general loss functions, such as those that arise from generalized linear models (Negahban et al. 2012, van de Geer 2008), as well as for certain types of nonconvex loss functions (Loh & Wainwright 2012).

3.2. Bounding the Statistical Error

In addition to a curvature condition, bounding the statistical error of a regularized M-estimator requires some measure of the "complexity" of the target parameter θ^*, as reflected by the regularizer. One way of doing so is by specifying a model subspace $\mathcal{M}$ and requiring that θ^* lie in the space (or is close to it). For instance, in the case of the group Lasso (Equation 6), a natural choice would be to fix a subset of groups $S \subseteq \mathcal{G}$ and then consider the subspace $\mathcal{M}(\mathcal{S}) = \{\theta \in \mathbb{R}^d | \theta_G = 0 \quad \text{for all } G \notin S\}$.

For a given model subspace $\mathcal{M}$, its complexity relative to the regularizer and the norm $\|\cdot\|$ used to measure the error can be measured in terms of the quantity

$$\Psi(\mathcal{M}) := \sup_{\theta \in \mathcal{M} \setminus \{0\}} \frac{\mathcal{R}(\theta)}{||\theta||}. \tag{21.}$$

For instance, in the context of the basic group Lasso (Equation 6), $||\theta||_{\mathcal{G}}/||\theta||_2 \leq \sqrt{|S|}$ for all $\theta \in \mathcal{M}(S)$, so that $\Psi(\mathcal{M}(\mathcal{S})) = \sqrt{|S|}$.

An additional condition is that the regularizer be decomposable with respect to $\mathcal{M}$. In the most stringent form of decomposability, the regularizer is required to satisfy the condition

$$\mathcal{R}(\alpha + \beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \quad \text{for all } \alpha \in \mathcal{M}, \quad \beta \in \mathcal{M}^{\perp}, \tag{22.}$$

where $\mathcal{M}^{\perp}$ denotes the orthogonal subspace. This form of decomposability is satisfied for the ℓ_1-penalty and group Lasso penalty. For the nuclear norm, a slightly relaxed form of decomposability holds (for further details on this and other aspects of decomposability, see Negahban et al. 2012).

Table 1 Examples of $(\mathcal{R}, \mathcal{R}^*)$ pairs[a]

Problem	Regularizer $\mathcal{R}$	Dual norm $\mathcal{R}^*$
Sparse vector	ℓ_1-norm	ℓ_∞-norm
	$\mathcal{R}(\theta) = \sum_{j=1}^{d} \vert\theta_j\vert$	$\mathcal{R}^*(\gamma) = \max_{j=1,\ldots,d} \vert\gamma_j\vert$
Group sparsity (nonoverlapping groups)	Group Lasso norm	Max. group norm
	$\mathcal{R}(\theta) = \sum_{G\in\mathcal{G}} \Vert\theta_G\Vert$	$\mathcal{R}^*(\gamma) = \max_{G\in\mathcal{G}} \Vert\gamma_G\Vert_*$
Group sparsity (overlapping groups)	Overlap group norm	Max. group norm
	$\mathcal{R}(\theta) = \inf'_{\substack{\theta=\sum_{G\in\mathcal{G}} w_G \\ w_G,\, G\in\mathcal{G}}} \left\{\sum_{G\in\mathcal{G}} \Vert w_G\Vert\right\}$	$\mathcal{R}^*(\gamma) = \max'_{G\in\mathcal{G}} \Vert\gamma_G\Vert_*$
Low-rank matrices	Nuclear norm	Operator norm
	$\mathcal{R}(\theta) = \sum_{j=1}^{d} \sigma_j(\Theta)$	$\mathcal{R}^*(\Gamma) = \vert\vert\vert\Gamma\vert\vert\vert_{\mathrm{op}} := \max_{j=1,\ldots,d} \sigma_j(\Gamma)$
Matrix decomposition	Additive decomposition norm	Max. norm
	$\mathcal{R}(\Theta) = \inf_{\Theta=A+B} \{\Vert A\Vert_1 + \omega \vert\vert\vert B\vert\vert\vert_{\mathrm{nuc}}\}$	$\mathcal{R}^*(\Gamma) = \max\{\Vert\Gamma\Vert_\infty, \omega^{-1}\vert\vert\vert\Gamma\vert\vert\vert_{\mathrm{op}}\}$
Sparse nonparametric regression $f = \sum_{j=1}^{d} f_j$	Sparse Hilbert norm	Max. dual norm
	$\mathcal{R}(f) = \sum_{j=1}^{d} \sqrt{\sum_{k=1}^{\infty} \frac{\theta_{jk}^2}{\mu_k}}$	$\mathcal{R}^*(g) = \max_{j=1,\ldots,d} \sqrt{\sum_{k=1}^{\infty} \mu_k \gamma_{jk}^2}$
	$f_j = \sum_{k=1}^{\infty} \theta_{jk}\phi_{jk}$	$g_j = \sum_{k=1}^{\infty} \gamma_{jk}\phi_j$

[a] For the group Lasso, the quantity $\| \cdot \|_*$ denotes the dual norm of $\| \cdot \|$. The choice of regularization parameter λ_n based on the dual norm bound (Equation 23) yields minimax-optimal rates in many settings.

The final question concerns how the regularization parameter λ_n in the M-estimator (Equation 3) should be chosen. The following result requires only that it be "large enough" to cancel out the effects of the effective noise. The noise can be measured in the terms of the score function—more specifically, the gradient $\nabla\mathcal{L}_n(\theta^*; Z_1^n)$ evaluated at the parameter θ^*. When θ^* is an unconstrained minimum of the population loss, then this score function is a zero-mean random vector, with fluctuations induced by having a finite number of samples. The resulting theorem involves conditioning on the "good" event $\mathbb{G}(\lambda_n; Z_1^n)$ such that the regularization parameter satisfies the dual norm bound, namely

$$\mathbb{G}(\lambda_n; Z_1^n) := \left\{\lambda_n \geq 2\, \mathcal{R}^*\left(\nabla\mathcal{L}_n(\theta^*; Z_1^n)\right)\right\}, \tag{23.}$$

where $\mathcal{R}^*(v) := \sup_{\mathcal{R}(u)\leq 1} \langle u, v\rangle$ is the dual norm defined by the regularizer $\mathcal{R}$ (for various examples, see **Table 1**). With these ingredients, we have the following result.

Theorem 1. For an $\mathcal{M}$-decomposable regularizer $\mathcal{R}$ with $\theta^* \in \mathcal{M}$, suppose that the empirical loss satisfies an RSC condition with parameters (κ, τ_n) and that the sample size is large enough so that $\tau_n \Psi^2(\mathcal{M}) < \frac{\kappa}{8}$. Conditioned on the event $\mathbb{G}(\lambda_n; Z_1^n)$, any global optimum $\hat{\theta}$ of the convex program given by Equation 3 satisfies the bound

$$||\hat{\theta} - \theta^*||^2 \leq \frac{9}{\kappa^2}\lambda_n^2 \Psi^2(\mathcal{M}). \tag{24.}$$

This result is a special case of a more general oracle inequality due to Negahban et al. (2012) that involves both estimation error (Equation 24) and an additional term of approximation error as well as a milder form of restricted strong convexity. Theorem 1 shows that, whenever the regularization parameter λ_n satisfies the dual norm bound (Equation 23), all optima of the family of convex programs (Equation 3) have desirable properties. To apply it in statistical settings, we must determine a choice of λ_n for which the dual norm bound (Equation 23) holds with high probability. This calculation depends on the structure of the score function defined by the loss function. Oracle inequalities can also be obtained under a slightly relaxed version of the decomposability condition (Equation 22) known as weak decomposability [for results of this type, see van de Geer (2012)]. Theorem 1 can also be used to recover a variety of error bounds for different types of M-estimators; let us consider a few cases here.

3.2.1. Group-structured sparsity. Consider a collection $\mathcal{G}$ of nonoverlapping groups, where the size of any group is at most $g_{\max}$. Suppose that θ^* is supported on a subset $\mathcal{S} \subseteq \mathcal{G}$ of groups, with the group Lasso (with the ℓ_2-norm within each coordinate) used as the regularizer. Suppose moreover that we draw i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ in which the covariates are zero-mean with sub-Gaussian tails, and conditioned on x_i, the response variable follows a generalized linear model (specified in terms of the inner product $\langle x_i, \theta^* \rangle$). Using the negative log-likelihood to define the empirical result, the choice $\lambda_n^2 \asymp \frac{g_{\max}}{n} + \frac{\log|\mathcal{G}|}{n}$ satisfies the conditions of Theorem 1. Therefore, under an appropriate form of restricted strong convexity for the group norm (Equation 6), the general error bound (Equation 24) implies that any group Lasso solution satisfies the bound

$$||\hat{\theta} - \theta^*||_2^2 \precsim \frac{\sigma^2}{\kappa^2}\left\{\frac{|\mathcal{S}|g_{\max}}{n} + \frac{|\mathcal{S}|\log|\mathcal{G}|}{n}\right\} \qquad 25.$$

with high probability. In the special case of $|\mathcal{G}| = d$ groups each of size $g_{\max} = 1$, the group Lasso reduces to the ordinary Lasso, and the bound (Equation 25) implies that $||\hat{\theta} - \theta^*||_2^2 \precsim \frac{\sigma^2}{\kappa^2}\frac{s\log d}{n}$ for any s-sparse vector θ^*. Lasso bounds of this type have been derived by various authors (e.g., Bickel et al. 2009, Bunea et al. 2007, van de Geer & Buhlmann 2009) and are minimax optimal (Raskutti et al. 2011). Bounds of the more general form (Equation 25) have also been derived by several authors (Huang & Zhang 2010, Lounici et al. 2011, Negahban et al. 2012) and are minimax optimal (Raskutti et al. 2012).

3.2.2. Low-rank matrices and nuclear norm. Consider a matrix $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ of rank r: A simple argument based on its SVD shows that it has slightly less than $r(d_1 + d_2)$ degrees of freedom. This quantity emerges naturally from Theorem 1 if it is applied to M-estimators of a rank r matrix using the nuclear norm as the regularizer. In particular, suppose that we observe n i.i.d. samples $\{(X_i, y_i)\}_{i=1}^n$ from the matrix regression model (Equation 20). For Gaussian noise $w_i \sim N(0, \sigma^2)$ and various choices of the observation matrices $\{X_i\}_{i=1}^n$, the dual norm bound (Equation 23) [where $\mathcal{R}^*(\Gamma) = |||\Gamma|||_{\text{op}}$ is the maximum singular value] is satisfied by choosing $\lambda_n \asymp \sigma\frac{d_1+d_2}{n}$. For a subspace $\mathcal{M}$ that consists of rank r matrices sharing the same singular vectors, we have $\Psi^2(\mathcal{M}) = r$, so that Theorem 1 implies a bound of the form

$$|||\hat{\Theta} - \Theta^*|||_{\text{F}}^2 \precsim \frac{\sigma^2}{\kappa^2}\frac{r(d_1 + d_2)}{n}. \qquad 26.$$

Bounds of this form have been derived for random linear projections (Candès & Plan 2011, Negahban & Wainwright 2011a, Oymak et al. 2012), for more general instances of matrix regression (Negahban & Wainwright 2011a, Rohde & Tsybakov 2011), for estimation of autoregressive processes (Negahban & Wainwright 2011a), and with an additional logarithmic factor for matrix completion (Koltchinskii et al. 2011, Negahban & Wainwright 2012).

3.2.3. Sparse additive nonparametric models. Arguments closely related to those underlying Theorem 1 are used to analyze the class of estimators (Equation 15) for the SpAM class

(Equation 13). Here there are two choices of regularization parameter, which requires a more sophisticated analysis than required by the dual norm bound (Equation 23). Ultimately, however, it leads to easily interpretable bounds on the error $(\hat{f} - f^*)$, as measured in the $L^2(\mathbb{P})$-norm. In particular, consider an unknown function f^* in the SpAM class (Equation 13), supported on s out of d coordinates, with $\mathcal{H}_j$ equal to a common univariate Hilbert space in all coordinates. Suppose that we observe samples of the form $y_i = f^*(x_i) + w_i$, where $w_i \sim N(0, \sigma^2)$ and the covariate vectors $x_i \in \mathbb{R}^d$ are sampled from a distribution with i.i.d. coordinates. In this setting, the estimator (Equation 15), with suitable choices of the regularization parameters, yields an estimate $\hat{f}$ such that

$$||\hat{f} - f^*||_2^2 \precsim \underbrace{\sigma^2 \frac{s \log d}{n}}_{\text{Cost of subset selection}} + \underbrace{s\delta_n^2}_{\text{Univariate estimation}} \qquad 27.$$

with high probability. The two terms in this bound have a very natural interpretation: The first corresponds to the cost of subset selection—that is, determining which of the s coordinates are active. The second term is s times the mean-squared error of estimating a single univariate function in the Hilbert space $\mathcal{H}$, a quantity that varies according to the smoothness of the univariate function classes. For instance, if the univariate functions are assumed to be twice differentiable, then the mean-squared error per coordinate scales as $\delta_n^2 \asymp (\sigma^2/n)^{4/5}$. Bounds of this form are established by Koltchinskii & Yuan (2010) and Raskutti et al. (2012).

3.3. Some Optimization Algorithms and Theory

Let us now turn to a brief discussion of some algorithms for solving optimization problems of the forms given in Equations 2 and 3. Optimization methods can be classified in terms of first versus second order, depending on whether they use only gradient-based information versus calculations of both first and second derivatives (Bertsekas 1995, Boyd & Vandenberghe 2004, Nesterov 2004). The convergence rates of second-order methods are usually faster with the caveat that each iteration is more expensive. First-order methods scale better to the large-scale problems that arise in high-dimensional statistics, and a great deal of recent work has studied such algorithms for regularized M-estimators (e.g., Bach et al. 2012; Bredies & Lorenz 2008; d'Aspremont et al. 2008; Duchi et al. 2008; Friedman et al. 2008; Fu 2001; Nesterov 2007, 2012; Tseng & Yun 2009; Wu & Lange 2008).

The projected gradient method is one of the simplest for solving the constrained problem (Equation 2). It is specified by a sequence $\{\alpha^t\}_{t=0}^{\infty}$ of positive step sizes and a starting point $\theta^0 \in \Omega$. For iterations $t = 0, 1, 2, \ldots$, given the current iterate θ^t, it generates the next iterate θ^{t+1} by taking a step in the direction of the negative gradient and then projecting the resulting vector $\theta^t - \alpha^t \nabla \mathcal{L}_n(\theta^t)$ back onto the constraint set.[3] Alternatively, it can be understood as forming a quadratic approximation to the loss around the current iterate and minimizing it over the constraint set:

$$\theta^{t+1} = \arg \min_{\{\theta \in \Omega | \mathcal{R}(\theta) \leq \rho\}} \left\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \theta - \theta^t \rangle + \frac{1}{2\alpha^t} ||\theta - \theta^t||_2^2 \right\}. \qquad 28.$$

For the Lagrangian M-estimator (Equation 3), a closely related method is the composite gradient algorithm (Nesterov 2007): It minimizes the sum of a quadratic approximation to the loss and the

[3] Throughout this section, for compactness in notation, we omit the dependence of the loss and its gradients on the data Z_1^n.

regularization penalty

$$\theta^{t+1} = \arg\min_{\theta\in\Omega} \left\{ \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t), \theta - \theta^t \rangle + \frac{1}{2\alpha^t} ||\theta - \theta^t||_2^2 + \lambda_n \mathcal{R}(\theta) \right\}. \qquad 29.$$

The efficiency of these methods depends on the complexity of evaluating the gradient term $\nabla\mathcal{L}_n(\theta^t)$ as well as on the complexity of performing the Euclidean projection onto the constraint set (Equation 28) or of solving the regularized quadratic problem (Equation 29). For the ℓ_1-norm, these updates can be performed by a soft-thresholding operation, and the resulting algorithm is known as iterative soft thresholding. The group Lasso update is similar, with coordinate-wise soft thresholding replaced by a block thresholding operation. For the nuclear norm, the update is also relatively simple, including an SVD calculation followed by soft thresholding of the singular values (for further discussion of these proximal updates for more complicated norms with overlapping groups and/or hierarchical structures, see Bach et al. 2012, section 3).

Another line of work has focused on the use of coordinate descent methods to solve regularized *M*-estimators (e.g., d'Aspremont et al. 2008, Friedman et al. 2008, Fu 2001, Nesterov 2012, Tseng 2001, Tseng & Yun 2009). These methods operate by choosing a single coordinate—or, more generally, a block of coordinates—to be updated within each iteration. In general, coordinate descent methods are not guaranteed to converge to a global optimum (Bertsekas 1995), but for regularizers that satisfy a certain separability condition (such as the ℓ_1- and group Lasso norms), convergence is always guaranteed (Tseng 2001). An attractive feature of block coordinate descent schemes is that each update can be substantially less expensive than updates such as Equation 28 or 29 that involve the full gradient. However, the convergence rates of coordinate descent methods are typically slower than those of proximal methods (Nesterov 2004, 2012), so there is a trade-off between the per-iteration complexity versus the total number of iterations.

Interestingly, the same conditions used to bound the statistical error of an *M*-estimator can be used to analyze the convergence rates of optimization algorithms. An optimization algorithm is said to exhibit sublinear convergence if there is a constant $c > 0$ such that $||\theta^t - \hat{\theta}||^2 \precsim (1/t)^c$ for iterations $t = 1, 2, \ldots$. Under convexity and Lipschitz conditions on the loss, both the projected gradient (Equation 28) and composite gradient (Equation 29) updates are guaranteed to converge at such a sublinear rate. A much faster type of convergence—indeed, the fastest type that can be expected from a first-order method in general (Nesterov 2004)—is linear or geometric convergence, for which there exists a contraction coefficient $\kappa \in (0, 1)$ such that $||\theta^t - \hat{\theta}||^2 \precsim \kappa^t ||\theta^0 - \hat{\theta}||^2$ for all $t = 1, 2, \ldots$. When the loss function $\mathcal{L}_n$ is both strongly convex and smooth, the algorithms given by Equations 28 and 29, with an appropriately chosen constant step size, are guaranteed to converge at a linear rate; smoothness here means that the Taylor series error (Equation 16) is upper bounded by a quadratic form.

As noted above for the high-dimensional setting, loss functions are not strongly convex. However, given that they often satisfy a restricted strong form of strong convexity (Equation 17), it is natural to wonder whether this can be leveraged to establish fast convergence rates. This intuition turns out to be correct: When the loss satisfies the restricted strong convexity condition (Equation 17) and an analogous upper bound, then first-order methods converge at a fast linear rate, at least up to the statistical precision of the problem. For instance, Agarwal et al. (2012a) showed that, for many statistical models, the projected gradient updates given by Equation 28 satisfy (with high probability) a bound of the form

$$\underbrace{||\theta^t - \hat{\theta}||^2}_{\text{Optimization error}} \leq \kappa^t ||\theta^0 - \hat{\theta}||^2 + o\left(\underbrace{||\hat{\theta} - \theta^*||^2}_{\text{Statistical error}}\right) \quad \text{for all iterations } t = 1, 2, \ldots, \qquad 30.$$

where $\kappa \in (0, 1)$ is a contraction factor that depends on the conditioning of the population-level problem. For instance, in the case of least-squares loss with random design, the bound in Equation 30 holds with $\kappa = 1 - \frac{1}{4}c(\Sigma)$, where $c(\Sigma)$ denotes the inverse condition number of the covariance matrix Σ of the covariates.

The bound given by Equation 30 guarantees fast linear convergence of the optimization error $\theta^t - \hat{\theta}$ up to the statistical precision of the problem, as given by the statistical error $\hat{\theta} - \theta^*$, where $\hat{\theta}$ is the global optimum of the regularized M-estimator and θ^* is the target parameter. **Figure 4*b*** illustrates the underlying geometry of the result. The optimization error $\Delta^t := \theta^t - \hat{\theta}$ decreases geometrically up to a certain tolerance, and the bound given by Equation 30 guarantees that this optimization tolerance is of lower order than the squared statistical error $||\hat{\theta} - \theta^*||^2$. Consequently, by taking $T \asymp \log(\kappa/\varepsilon^2)$ iterations, we are guaranteed that θ^T has statistical error of the same order as the global optimum. This result, though somewhat unusual from an optimization-theoretic perspective, is precisely the type of guarantee that is relevant to the statistician, for whom θ^* is the quantity of ultimate interest.

4. DISCUSSION AND FUTURE DIRECTIONS

Recent years have witnessed a great deal of progress on convex M-estimators with structured regularizers for high-dimensional problems. They are used in a wide variety of applications and are equipped with attractive statistical guarantees. In addition, the underlying optimization problems can be solved efficiently by various iterative algorithms. As outlined in this survey, relatively well-understood aspects include modeling questions, such as which regularizers are well suited to certain types of structure, and the properties inherited by the solutions. Basic aspects of statistical theory are in place, including nonasymptotic bounds on estimation error and the minimum sample size required for correct model selection. In many but not all cases, the achievable results guaranteed by convex M-estimators achieve minimax lower bounds. Finally, there has been a great deal of work on fast algorithms, including bounds on convergence rates.

Despite this encouraging progress, there remain a variety of interesting open questions. To date, the bulk of statistical theory has focused on bounding the estimation error. For inferential purposes, it is also important to obtain confidence values associated with these estimates, and an emerging line of research (e.g., Javanmard & Montanari 2013, van de Geer et al. 2013, Zhang & Zhang 2011) is developing this type of understanding. Other interesting questions concern the trade-offs between convex and nonconvex methods. For many problems, estimators based on convex programs are minimax optimal up to constant factors, but compared with nonconvex methods, they often require more stringent conditions. As an example, methods based on ℓ_1-regularization achieve minimax-optimal rates for vector estimation, but their success requires more severe conditions on the design matrix (Raskutti et al. 2011) than are required by nonconvex methods such as ℓ_0-based regularization. It is an open question to what extent these gaps can be closed by polynomial-time estimators. In other cases, there are known gaps between the best-known performance of polynomial-time methods and that of optimal but computationally intractable methods. Examples include variable selection and testing in sparse principal component analysis (Amini & Wainwright 2009, Berthet & Rigollet 2013) as well as denoising of sparse and low-rank matrices (Oymak et al. 2012). Finally, a variety of open questions are associated with high-dimensional nonparametric and semiparametric models.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

This work was partially funded by the National Science Foundation DMS grant 1107000. Thanks to the anonymous reviewer for careful reading and insightful comments.

LITERATURE CITED

Agarwal A, Negahban S, Wainwright MJ. 2012a. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Ann. Stat.* 40(5):2452–82

Agarwal A, Negahban S, Wainwright MJ. 2012b. Noisy matrix decomposition via convex relaxation: optimal rates in high dimensions. *Ann. Stat.* 40(2):1171–97

Amini AA, Wainwright MJ. 2009. High-dimensional analysis of semidefinite relaxations for sparse principal component analysis. *Ann. Stat.* 37(5B):2877–921

Bach F. 2008. Consistency of trace norm minimization. *J. Mach. Learn. Res.* 9:1019–48

Bach F, Jenatton R, Mairal J, Obozinski G. 2012. Optimization with sparsity-inducing penalties. *Found. Trends Mach. Learn.* 4(1):1–106

Baraniuk RG, Cevher V, Duarte MF, Hegde C. 2010. Model-based compressive sensing. *IEEE Trans. Inf. Theory* 56(4):1982–2001

Berthet Q, Rigollet P. 2013. *Computational lower bounds for sparse PCA*. Tech. Rep., Princeton Univ., Princeton, NJ. **http://arxiv1304.0828**

Bertsekas DP. 1995. *Nonlinear Programming*. Belmont, MA: Athena Sci.

Bickel P, Li B. 2006. Regularization in statistics. *TEST* 15(2):271–344

Bickel P, Ritov Y, Tsybakov A. 2009. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat.* 37(4):1705–32

Boyd S, Vandenberghe L. 2004. *Convex Optimization*. Cambridge, UK: Cambridge Univ. Press

Bredies K, Lorenz DA. 2008. Linear convergence of iterative soft thresholding. *J. Fourier Anal. Appl.* 14:813–37

Bühlmann P, van de Geer S. 2011. *Statistics for High-Dimensional Data*. New York: Springer

Bunea F, Tsybakov A, Wegkamp M. 2007. Sparsity oracle inequalities for the Lasso. *Electron. J. Stat.* 169–94

Cai T, Liu W, Luo X. 2011. A constrained ℓ_1-minimization approach to sparse precision matrix estimation. *J. Am. Stat. Assoc.* 106:594–607

Cai TT, Liu W, Zhou HH. 2012. *Estimating sparse precision matrices: optimal rates of convergence and adaptive estimation*. Tech. Rep., Wharton Sch., Univ. Pa., Pennsylvania, PA. **http://arxiv1212.2882**

Candès EJ, Li X, Ma Y, Wright J. 2011. Robust principal component analysis? *J. ACM* 58:11

Candès EJ, Plan Y. 2011. Tight oracle bounds for low-rank matrix recovery from a minimal number of random measurements. *IEEE Trans. Inf. Theory* 57(4):2342–59

Candès EJ, Recht B. 2009. Exact matrix completion via convex optimization. *Found. Comput. Math.* 9(6):717–72

Candès EJ, Tao T. 2007. The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.* 35(6):2313–51

Chandrasekaran V, Parrilo PA, Willsky AS. 2012. Latent variable graphical model selection via convex optimization. *Ann. Stat.* 40(4):1935–67

Chandrasekaran V, Sanghavi S, Parrilo PA, Willsky AS. 2011. Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optim.* 21:572–96

Chen S, Donoho DL, Saunders MA. 1998. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20(1):33–61

Cohen A, Dahmen W, DeVore R. 2009. Compressed sensing and best k-term approximation. *J. Am. Math. Soc.* 22(1):211–31

d'Aspremont A, Banerjee O, El Ghaoui L. 2008. First-order methods for sparse covariance selection. *SIAM J. Matrix Anal. Appl.* 30(1):55–66

Donoho DL. 2006. Compressed sensing. *IEEE Trans. Inf. Theory* 52(4):1289–306

Donoho DL, Tanner JM. 2008. Counting faces of randomly-projected polytopes when the projection radically lowers dimension. *J. Am. Math. Soc.* 22:1–53

Duchi J, Shalev-Shwartz S, Singer Y, Chandra T. 2008. *Efficient projections onto the ℓ_1 -ball for learning in high dimensions.* Presented at Int. Conf. Mach. Learn., 25th, Helsinki, Finland

Fan J, Li R. 2001. Variable selection via non-concave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* 96(456):1348–60

Fazel M. 2002. *Matrix rank minimization with applications*. PhD Thesis, Stanford Univ., Stanford, CA. **http://faculty.washington.edu/mfazel/thesis-final.pdf**

Friedman J, Hastie T, Tibshirani R. 2008. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9:432–41

Fu WJ. 2001. Penalized regression: the bridge versus the Lasso. *J. Comput. Graph. Stat.* 7(3):397–416

Greenshtein E, Ritov Y. 2004. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli* 10:971–88

Gross D. 2011. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inf. Theory* 57(3):1548–66

Gyorfi L, Kohler M, Krzyzak A, Walk H. 2002. *A Distribution-Free Theory of Nonparametric Regression*. New York: Springer

Hoerl AE, Kennard RW. 1970. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics* 12:55–67

Hsu D, Kakade SM, Zhang T. 2011. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Inf. Theory* 57(11):7221–34

Huang J, Zhang T. 2010. The benefit of group sparsity. *Ann. Stat.* 38(4):1978–2004

Jacob L, Obozinski G, Vert JP. 2009. *Group Lasso with overlap and graph Lasso*. Presented at the 26th Int. Conf. Mach. Learn., Montreal, Canada

Jalali A, Ravikumar P, Sanghavi SS, Ruan C. 2010. A dirty model for multi-task learning. *Adv. Neural Inf. Process. Syst.* 23:964–72

Javanmard A, Montanari A. 2013. *Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory*. Tech. Rep., Stanford Univ., Stanford, CA. **http://arxiv1301.4240**

Kim Y, Kim J, Kim Y. 2006. Blockwise sparse regression. *Stat. Sin.* 16(2):375–90

Koltchinskii V, Lounici K, Tsybakov AB. 2011. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Ann. Stat.* 39:2302–29

Koltchinskii V, Yuan M. 2008. *Sparse recovery in large ensembles of kernel machines*. Presented at Annu. Conf. Learn. Theory, 21st, Helsinki, Finland

Koltchinskii V, Yuan M. 2010. Sparsity in multiple kernel learning. *Ann. Stat.* 38:3660–95

Lam C, Fan J. 2009. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat.* 37:4254–78

Lin Y, Zhang HH. 2006. Component selection and smoothing in multivariate nonparametric regression. *Ann. Stat.* 34:2272–97

Liu H, Lafferty J, Wasserman L. 2009. The nonparanormal: semiparametric estimation of high-dimensional undirected graphs. *J. Mach. Learn. Res.* 10:1–37

Loh P, Wainwright MJ. 2012. High-dimensional regression with noisy and missing data: provable guarantees with non-convexity. *Ann. Stat.* 40(3):1637–64

Lounici K, Pontil M, Tsybakov AB, van de Geer S. 2011. Oracle inequalities and optimal inference under group sparsity. *Ann. Stat.* 39(4):2164–204

Mazumder R, Hastie T, Tibshirani R. 2010. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.* 11:2287–322

Meier L, van de Geer S, Buhlmann P. 2009. High-dimensional additive modeling. *Ann. Stat.* 37:3779–821

Meinshausen N, Bühlmann P. 2006. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat.* 34:1436–62

Micchelli CA, Morales JM, Pontil M. 2013. Regularizers for structured sparsity. *Adv. Comput. Math.* 38:455–89

Negahban S, Ravikumar P, Wainwright MJ, Yu B. 2012. A unified framework for high-dimensional analysis of *M*-estimators with decomposable regularizers. *Stat. Sci.* 27(4):538–57

Negahban S, Wainwright MJ. 2011a. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Ann. Stat.* 39(2):1069–97

Negahban S, Wainwright MJ. 2011b. Simultaneous support recovery in high-dimensional regression: benefits and perils of $\ell_{1,\infty}$-regularization. *IEEE Trans. Inf. Theory* 57(6):3841–63

Negahban S, Wainwright MJ. 2012. Restricted strong convexity and (weighted) matrix completion: optimal bounds with noise. *J. Mach. Learn. Res.* 13:1665–97

Nesterov Y. 2004. *Introductory Lectures on Convex Optimization*. New York: Kluwer Acad.

Nesterov Y. 2007. *Gradient methods for minimizing composite objective function*. Tech. Rep. 76, Cent. Oper. Res. Econom., Catholic Univ., Louvain, Belg.

Nesterov Y. 2012. Efficiency of coordinate descent methods on huge-scale optimization problems. *SIAM J. Optim.* 22(2):341–62

Obozinski G, Wainwright MJ, Jordan MI. 2011. Union support recovery in high-dimensional multivariate regression. *Ann. Stat.* 39(1):1–47

Oymak S, Jalali A, Fazel M, Eldar YC, Hassibi B. 2012. *Simultaneously structured models with applications to sparse and low-rank matrices*. Tech. Rep., Calif. Inst. Technol., Pasadena, CA. **http://arxiv1212.3753**

Raskutti G, Wainwright MJ, Yu B. 2010. Restricted eigenvalue conditions for correlated Gaussian designs. *J. Mach. Learn. Res.* 11:2241–59

Raskutti G, Wainwright MJ, Yu B. 2011. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory* 57(10):6976–94

Raskutti G, Wainwright MJ, Yu B. 2012. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *J. Mach. Learn. Res.* 12:389–427

Ravikumar P, Liu H, Lafferty J, Wasserman L. 2009. SpAM: sparse additive models. *J. R. Stat. Soc. Ser. B* 71(5):1009–30

Ravikumar P, Wainwright MJ, Raskutti G, Yu B. 2011. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat.* 5:935–80

Recht B. 2011. A simpler approach to matrix completion. *J. Mach. Learn. Res.* 12:3413–30

Recht B, Fazel M, Parrilo P. 2010. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev.* 52(3):471–501

Rohde A, Tsybakov A. 2011. Estimation of high-dimensional low-rank matrices. *Ann. Stat.* 39(2):887–930

Rothman AJ, Bickel PJ, Levina E, Zhu J. 2008. Sparse permutation invariant covariance estimation. *Electron. J. Stat.* 2:494–515

Rudelson M, Zhou S. 2012. Reconstruction from anisotropic random measurements. *IEEE Trans. Inf. Theory* 59:3434–47

Srebro N, Alon N, Jaakkola TS. 2005. *Generalization error bounds for collaborative prediction with low-rank matrices*. Presented at Neural Inf. Proc. Syst., 17th, Vancouver

Srebro N, Rennie J, Jaakkola TS. 2004. *Maximum-margin matrix factorization*. Presented at Neural Inf. Proc. Syst., 17th, Vancouver

Stojnic M, Parvaresh F, Hassibi B. 2009. On the reconstruction of block-sparse signals with an optimal number of measurements. *IEEE Trans. Signal Process.* 57(8):3075–85

Stone CJ. 1982. Optimal global rates of convergence for non-parametric regression. *Ann. Stat.* 10(4):1040–53

Stone CJ. 1985. Additive regression and other non-parametric models. *Ann. Stat.* 13(2):689–705

Tibshirani R. 1996. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* 58(1):267–88

Tikhonov AN. 1943. On the stability of inverse problems. *C. R. (Doklady) Acad. Sci. SSSR* 39:176–79

Tropp JA. 2006. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Trans. Inf. Theory* 52(3):1030–51

Tropp JA, Gilbert AC, Strauss MJ. 2006. Algorithms for simultaneous sparse approximation. Part I: greedy pursuit. *Signal Process.* 86:572–88

Tseng P. 2001. Convergence of block coordinate descent method for nondifferentiable maximization. *J. Opt. Theory Appl.* 109(3):474–94

Tseng P, Yun S. 2009. A block-coordinate gradient descent method for linearly constrained nonsmooth separable optimization. *J. Optim. Theory Appl.* 140:513–35

Turlach B, Venables WN, Wright SJ. 2005. Simultaneous variable selection. *Technometrics* 27:349–63

van de Geer S. 2000. *Empirical Processes in M-Estimation*. Cambridge, UK: Cambridge Univ. Press

van de Geer S. 2008. High-dimensional generalized linear models and the Lasso. *Ann. Stat.* 36:614–45

van de Geer S. 2012. *Weakly decomposable regularization penalties and structured sparsity*. Tech. Rep., ETH Zurich, Switz. **http://arxiv1204.4813v2**

van de Geer S, Buhlmann P. 2009. On the conditions used to prove oracle results for the Lasso. *Electron. J. Stat.* 3:1360–92

van de Geer S, Buhlmann P, Ritov Y. 2013. *On asymptotically optimal confidence regions and tests for high-dimensional models*. Tech. Rep., ETH Zurich, Switz. **http://arxiv1303.0518**

Wainwright MJ. 2009. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory* 55:2183–202

Wu TT, Lange K. 2008. Coordinate descent algorithms for Lasso-penalized regression. *Ann. Appl. Stat.* 2(1):224–44

Xu H, Caramanis C, Sanghavi S. 2012. Robust PCA via outlier pursuit. *IEEE Trans. Inf. Theory* 58(5):3047–64

Yuan M. 2010. High-dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.* 11:2261–86

Yuan M, Lin Y. 2006. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B* 68:49–67

Zhang CH, Huang J. 2008. The sparsity and bias of the Lasso selection in high-dimensional linear regression. *Ann. Stat.* 36(4):1567–94

Zhang CH, Zhang SS. 2011. *Confidence intervals for low-dimensional parameters with high-dimensional data*. Tech. Rep., Rutgers Univ., New Brunswick, NJ. **http://arxiv1110.2563**

Zhang CH, Zhang T. 2012. A general theory of concave regularization for high-dimensional sparse estimation problems. *Stat. Sci.* 27(4):576–93

Zhao P, Rocha G, Yu B. 2009. Grouped and hierarchical model selection through composite absolute penalties. *Ann. Stat.* 37(6A):3468–97

Zhao P, Yu B. 2006. On model selection consistency of Lasso. *J. Mach. Learn. Res.* 7:2541–67

Zhou S, Lafferty J, Wasserman L. 2008. Time-varying undirected graphs. Presented at Annu. Conf. Learn. Theory, 21st, Helsinki

Annual Review of Statistics and Its Application

Volume 1, 2014

Contents