Submitted to the Annals of Statistics

FAST GLOBAL CONVERGENCE OF GRADIENT METHODS FOR HIGH-DIMENSIONAL STATISTICAL RECOVERY

By Alekh Agarwal[⋆] and Sahand Negahban[‡]
and Martin J. Wainwright[†]

UC Berkeley, Department of EECS[⋆,†] *and Statistics*[†]
MIT, Department of EECS[‡]

Many statistical M-estimators are based on convex optimization problems formed by the combination of a data-dependent loss function with a norm-based regularizer. We analyze the convergence rates of projected gradient and composite gradient methods for solving such problems, working within a high-dimensional framework that allows the ambient dimension d to grow with (and possibly exceed) the sample size n. Our theory identifies conditions under which projected gradient descent enjoys globally linear convergence up to the *statistical precision* of the model, meaning the typical distance between the true unknown parameter θ^* and an optimal solution $\widehat{\theta}$. By establishing these conditions with high probability for numerous statistical models, our analysis applies to a wide range of M-estimators, including sparse linear regression using Lasso; group Lasso for block sparsity; log-linear models with regularization; low-rank matrix recovery using nuclear norm regularization; and matrix decomposition using a combination of the nuclear and ℓ_1 norms. Overall, our analysis reveals interesting connections between statistical and computational efficiency in high-dimensional estimation.

1. Introduction. High-dimensional data sets present challenges that are both statistical and computational in nature. On the statistical side, recent years have witnessed a flurry of results on convergence rates for various estimators under high-dimensional scaling, allowing for the possibility that the problem dimension d exceeds the sample size n. These results typically involve some assumption regarding the structure of the parameter space, such as sparse vectors, structured covariance matrices, or low-rank matrices, as well as some regularity of the data-generating process. On the computational side, many estimators for statistical recovery are based on solving convex programs. Examples of such M-estimators include ℓ_1-regularized quadratic programs (Lasso) for sparse linear regression

AMS 2000 subject classifications: Primary 62F30, 62F30; secondary 62H12
Keywords and phrases: high-dimensional inference, convex optimization, regularized M-estimation

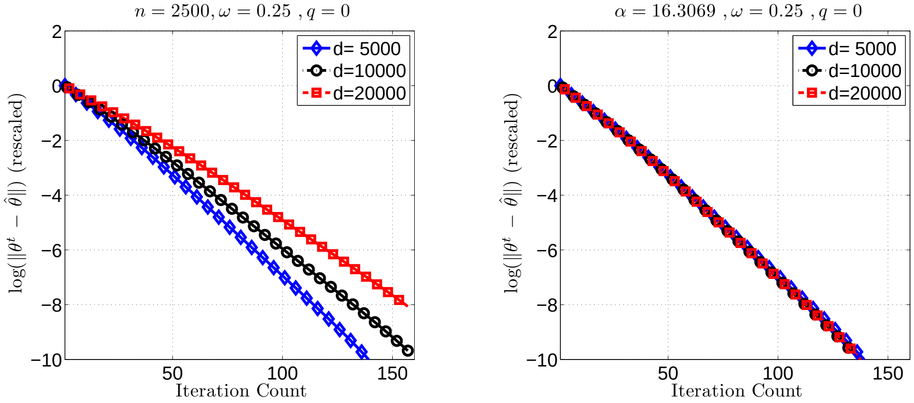
(e.g. [40, 13, 44, 26, 7]), second-order cone programs (SOCP) for the group Lasso (e.g., [45, 24, 19]), and SDP relaxations for various problems, including sparse PCA and low-rank matrix estimation (e.g., [11, 39, 3, 37, 28, 36]).

Many of these programs are instances of convex conic programs, and so can (in principle) be solved to ϵ-accuracy in polynomial time using interior point methods, and other standard methods from convex programming (e.g., see the books [6, 8]). However, the complexity of such quasi-Newton methods can be prohibitively expensive for the very large-scale problems that arise from high-dimensional data sets. Accordingly, recent years have witnessed a renewed interest in simpler first-order methods, among them the methods of projected gradient descent and mirror descent. Several authors (e.g., [5, 20, 4]) have used variants of Nesterov's accelerated gradient method [31] to obtain algorithms for high-dimensional statistical problems with a sublinear rate of convergence. Note that an optimization algorithm, generating a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$, is said to exhibit *sublinear convergence* to an optimum $\widehat{\theta}$ if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at the rate $1/t^{\kappa}$, for some exponent $\kappa > 0$ and norm $\|\cdot\|$. It is known that this is the best possible convergence rate for gradient descent-type methods for convex programs under only Lipschitz conditions [30].

It is known that much faster global rates—in particular, a linear or geometric rate—can be achieved if global regularity conditions like strong convexity and smoothness are imposed [30]. An optimization algorithm is said to exhibit *linear or geometric* convergence if the optimization error $\|\theta^t - \widehat{\theta}\|$ decays at a rate κ^t, for some contraction coefficient $\kappa \in (0, 1)$. Note that such convergence is exponentially faster than sub-linear convergence. For certain classes of problems involving polyhedral constraints and global smoothness, Tseng and Luo [25] have established geometric convergence. However, a challenging aspect of statistical estimation in high dimensions is that the underlying optimization problems can never be strongly convex in a global sense when $d > n$ (since the $d \times d$ Hessian matrix is rank-deficient), and global smoothness conditions cannot hold when $d/n \to +\infty$. Some more recent work has exploited structure specific to the optimization problems that arise in statistical settings. For the special case of sparse linear regression with random isotropic designs (also referred to as compressed sensing), some authors have established local linear convergence, meaning guarantees that apply once the iterates are close enough to the optimum [9, 17]. Also in the setting of compressed sensing, Tropp and Gilbert [41] studied finite convergence of greedy algorithms, while Garg and Khandekar [16] provide results for a thresholded gradient algorithm. In both of these results, the convergence happens up to a tolerance of the order of the noise variance, which is substantially larger than the true statistical precision of the problem.

The focus of this paper is the convergence rate of two simple gradient-based algorithms for solving optimization problems that underlie regularized M-estimators. For a constrained problem with a differentiable objective function, the projected gradient method generates a sequence of iterates $\{\theta^t\}_{t=0}^{\infty}$ by taking a step in the negative gradient direction, and then projecting the result onto the constraint set. The composite gradient method of Nesterov [31] is well-suited to solving regularized problems formed by the sum of a differentiable and a non-differentiable component.

The main contribution of this paper is to establish a form of global geometric convergence for these algorithms that holds for a broad class of high-dimensional statistical problems. In order to provide intuition for this guarantee, Figure 1 shows the performance of projected gradient descent for Lasso problems (ℓ_1-constrained least-squares), each one based on a fixed sample size $n = 2500$ and varying dimensions $d \in \{5000, 10000, 20000\}$. In panel (a), we have plotted the logarithm of the optimization error, measured in terms of the Euclidean norm $\|\theta^t - \widehat{\theta}\|$ between θ^t and an optimal solution $\widehat{\theta}$, versus the iteration number t. Note that all curves are linear (on this logarithmic scale), revealing the geometric convergence predicted by our theory.



(a) (b)

Fig 1. Convergence rates of projected gradient descent in application to Lasso (ℓ_1-constrained least-squares). Each panel shows the log optimization error $\log\|\theta^t - \widehat{\theta}\|$ versus the iteration number t. Panel (a) shows three curves, corresponding to dimensions $d \in \{5000, 10000, 20000\}$, sparsity $s = \lceil\sqrt{d}\rceil$, and all with the same sample size $n = 2500$. All cases show geometric convergence, but the rate for larger problems becomes progressively slower. (b) For an appropriately rescaled sample size ($\alpha = \frac{n}{s \log d}$), all three convergence rates should be roughly the same, as predicted by the theory.

Moreover, the results in panel (a) exhibit an interesting property: the convergence rate is *dimension-dependent*, meaning that for a fixed sample size, projected gradient descent converges more slowly for a large problem than a smaller problem. This phenomenon reflects the natural intuition that larger problems are "harder" than smaller problems. A notable aspect of our theory is that it makes a quantitative prediction regarding the extent to which a larger problem is harder than a smaller one. In particular, our convergence rates suggest that if the sample size n is re-scaled according to the dimension d and also other model parameters such as sparsity, then convergence rates should be roughly similar. Panel (b) confirms this prediction: when the sample size is rescaled according to our theory (in particular, see Corollary 2 in Section 3.2), then all three curves lie essentially on top of another.

Although high-dimensional optimization problems are typically neither strongly convex nor smooth, this paper shows that it is fruitful to consider suitably restricted notions of strong convexity and smoothness. Our notion of restricted strong convexity (RSC) is related to but slightly different than that of Negahban et al. [27] for establishing statistical consistency. We also introduce a related notion of restricted smoothness (RSM), not needed for proving statistical rates but essential in the setting of optimization. Our analysis consists of two parts. We first show that for optimization problems underlying many regularized M-estimators, RSC/RSM conditions are sufficient to guarantee global linear convergence of projected gradient descent. Our second contribution is to prove that for the iterates generated by our methods, these RSC/RSM assumptions do hold with high probability for numerous statistical models, among them sparse linear models, models with group sparsity, and various matrix estimation problems, including matrix completion and matrix decomposition.

An interesting aspect of our results is that the geometric convergence is not guaranteed to an arbitrary precision, but only to an accuracy related to *statistical precision* of the problem. For a given norm $\|\cdot\|$, the statistical precision is given by the mean-squared error $\mathbb{E}[\|\widehat{\theta}-\theta^*\|^2]$ between the true parameter θ^* and the solution $\widehat{\theta}$ of the optimization problem. Our analysis guarantees geometric convergence to a parameter θ such that

$$\|\theta-\theta^*\| = \|\widehat{\theta}-\theta^*\| + o(\|\widehat{\theta}-\theta^*\|),$$

which is the best we can hope for statistically, ignoring lower order terms. Overall, our results reveal an interesting connection between the statistical and computational properties of M-estimators—that is, the properties of the underlying statistical model that make it favorable for estimation also render it more amenable to optimization procedures.

The remainder of this paper is organized as follows. We begin in Section 2 with our setup and the necessary background. Section 3 is devoted to the statement of our main results and various corollaries. In Section 4, we provide a number of empirical results that confirm the sharpness of our theory. Proofs of our results have been provided in the supplementary material [2].

2. Background and problem formulation.

In this section, we begin by describing the class of regularized M-estimators to which our analysis applies, as well as the optimization algorithms that we analyze. Finally, we introduce some important notions that underlie our analysis, including the notions of a decomposable regularization, and the properties of restricted strong convexity and smoothness.

2.1. *Loss functions, regularization and gradient-based methods.*

Given a random variable $Z \sim \mathbb{P}$ taking values in some set $\mathcal{Z}$, let $Z_1^n = \{Z_1, \ldots, Z_n\}$ be a sample of n observations. Assuming that $\mathbb{P}$ lies within some indexed family $\{\mathbb{P}_\theta, \theta \in \Omega\}$, the goal is to recover an estimate of the unknown true parameter $\theta^* \in \Omega$ generating the data. Here Ω is some subset of $\mathbb{R}^d$, where d is the *ambient dimension* of the problem. In order to measure the "fit" of any $\theta \in \Omega$ to a given data set Z_1^n, we introduce a loss function $\mathcal{L}_n : \Omega \times \mathcal{Z}^n \to \mathbb{R}_+$. By construction, for any given n-sample data set $Z_1^n \in \mathcal{Z}^n$, the loss function assigns a cost $\mathcal{L}_n(\theta; Z_1^n) \geq 0$ to the parameter $\theta \in \Omega$. In many applications, the loss function has a separable structure across the data set, meaning that $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n}\sum_{i=1}^n \ell(\theta; Z_i)$ where $\ell : \Omega \times \mathcal{Z} :\to \mathbb{R}_+$ is the loss function associated with a single data point.

Of primary interest in this paper are estimation problems that are underdetermined, meaning that the sample size n is smaller than the ambient dimension d. In such settings, without further restrictions on the parameter space Ω, there are various impossibility theorems, asserting that consistent estimates of the unknown parameter θ^* cannot be obtained. For this reason, it is necessary to assume that the unknown parameter θ^* either lies within a smaller subset of Ω, or is well-approximated by some member of such a subset. In order to incorporate these types of structural constraints, we introduce a *regularizer* $\mathcal{R} : \Omega \to \mathbb{R}_+$ over the parameter space. Given a user-defined radius $\rho > 0$, our analysis applies to the *constrained M-estimator*

$$\widehat{\theta}_\rho \in \arg\min_{\mathcal{R}(\theta) \leq \rho} \big\{\mathcal{L}_n(\theta; Z_1^n)\big\}, \tag{1}$$

as well as to the *regularized M-estimator*

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\mathcal{R}(\theta) \leq \bar{\rho}} \big\{\underbrace{\mathcal{L}_n(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta)}_{\phi_n(\theta)}\big\}, \tag{2}$$

where the regularization weight $\lambda_n > 0$ is user-defined. Note that the radii ρ and $\bar{\rho}$ may be different in general. Throughout this paper, we impose the following two conditions:

(a) for any data set Z_1^n, the function $\mathcal{L}_n(\cdot; Z_1^n)$ is convex and differentiable over Ω, and
(b) the regularizer $\mathcal{R}$ is a norm.

These conditions ensure that the overall problem is convex, so that by Lagrangian duality, the optimization problems (1) and (2) are equivalent. However, as our analysis will show, solving one or the other can be computationally more preferable depending upon the assumptions made. When the radius ρ or the regularization parameter λ_n is clear from the context, we will drop the subscript on $\widehat{\theta}$ to ease the notation. Similarly, we frequently adopt the shorthand $\mathcal{L}_n(\theta)$. Procedures based on optimization problems of either form are known as M-estimators in the statistics literature.

The focus of this paper is on two simple algorithms for solving the above optimization problems. The method of *projected gradient descent* applies naturally to the constrained problem (1), whereas the *composite gradient descent* method due to Nesterov [31] is suitable for solving the regularized problem (2). Each routine generates a sequence $\{\theta^t\}_{t=0}^{\infty}$ of iterates by first initializing to some parameter $\theta^0 \in \Omega$, and then for $t = 0, 1, 2, \ldots$, applying the recursive update

$$\theta^{t+1} = \arg \min_{\theta \in \mathbb{B}_{\mathcal{R}}(\rho)} \big\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t),\, \theta - \theta^t \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|^2 \big\}, \tag{3}$$

in the case of projected gradient descent, or the update

$$\theta^{t+1} = \arg \min_{\theta \in \mathbb{B}_{\mathcal{R}}(\bar{\rho})} \big\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t),\, \theta - \theta^t \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|^2 + \lambda_n \mathcal{R}(\theta) \big\}, \tag{4}$$

for the composite gradient method. Note that the only difference between the two updates is the addition of the regularization term in the objective. These updates have a natural intuition: the next iterate θ^{t+1} is obtained by constrained minimization of a first-order approximation to the loss function, combined with a smoothing term that controls how far one moves from the current iterate in terms of Euclidean norm. Moreover, it is easily seen that the update (3) is equivalent to

$$\theta^{t+1} = \Pi\Big(\theta^t - \frac{1}{\gamma_u} \nabla \mathcal{L}_n(\theta^t) \Big), \tag{5}$$

where $\Pi \equiv \Pi_{\mathbb{B}_{\mathcal{R}}(\rho)}$ denotes Euclidean projection onto the regularizer norm ball $\mathbb{B}_{\mathcal{R}}(\rho) := \{\theta \in \Omega \mid \mathcal{R}(\theta) \leq \rho\}$ of radius ρ. In this formulation, we see that the algorithm takes a step in the negative gradient direction, using the quantity $1/\gamma_u$ as stepsize parameter, and then projects the resulting vector onto the constraint set. The update (4) takes an analogous form, however, the projection will depend on both λ_n and γ_u. As will be illustrated in the examples to follow, for many problems, the updates (3) and (4), or equivalently (5), have a very simple solution. For instance, in the case of ℓ_1-regularization, they are easily computed by an appropriate form of soft-thresholding.

2.2. *Restricted strong convexity and smoothness.* In this section, we define the conditions on the loss function and regularizer that underlie our analysis. Global smoothness and strong convexity assumptions play an important role in the classical analysis of optimization algorithms [6, 8, 30]. In application to a differentiable loss function $\mathcal{L}_n$, both of these properties are defined in terms of a first-order Taylor series expansion around a vector θ' in the direction of θ—namely, the quantity

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') := \mathcal{L}_n(\theta) - \mathcal{L}_n(\theta') - \langle \nabla \mathcal{L}_n(\theta'),\, \theta - \theta' \rangle. \tag{6}$$

By the assumed convexity of $\mathcal{L}_n$, this error is always non-negative, and global strong convexity is equivalent to imposing a stronger condition, namely that for some parameter $\gamma_\ell > 0$, the first-order Taylor error $\mathcal{T}_{\mathcal{L}}(\theta; \theta')$ is lower bounded by a quadratic term $\frac{\gamma_\ell}{2} \|\theta - \theta'\|^2$ for all $\theta, \theta' \in \Omega$. Global smoothness is defined in a similar way, by imposing a quadratic upper bound on the Taylor error. It is known that under global smoothness and strong convexity assumptions, the method of projected gradient descent (3) enjoys a *globally geometric convergence rate*, meaning that there is some $\kappa \in (0,1)$ such that[1]

$$\|\theta^t - \widehat{\theta}\|^2 \lesssim \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{7}$$

We refer the reader to Bertsekas [6, Prop. 1.2.3, p. 145], or Nesterov [30, Thm. 2.2.8, p. 88] for such results on projected gradient descent, and to Nesterov [31] for related results on composite gradient descent.

Unfortunately, in the high-dimensional setting ($d > n$), it is usually impossible to guarantee strong convexity of the problem (1) in a global sense. For instance, when the data is drawn i.i.d., the loss function consists of a sum of n terms. If the loss is twice differentiable, the resulting $d \times d$ Hessian matrix $\nabla^2 \mathcal{L}(\theta; Z_1^n)$ is often a sum of n matrices each with rank one,

[1] In this statement (and throughout the paper), we use $\lesssim$ to mean an inequality that holds with some universal constant c, independent of the problem parameters.

so that the Hessian is rank-degenerate when $n < d$. However, as we show in this paper, in order to obtain fast convergence rates for the optimization method (3), it is sufficient that (a) the objective is strongly convex and smooth in a restricted set of directions, and (b) the algorithm approaches the optimum $\widehat{\theta}$ only along these directions. Let us now formalize these ideas.

Definition 1 (**Restricted strong convexity (RSC)**). *The loss function $\mathcal{L}_n$ satisfies restricted strong convexity with respect to $\mathcal{R}$ and with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ over the set Ω' if*

$$\mathcal{T}_{\mathcal{L}}(\theta;\theta') \geq \frac{\gamma_\ell}{2}\,\|\theta-\theta'\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta-\theta') \qquad \text{for all } \theta,\theta' \in \Omega'. \tag{8}$$

We refer to the quantity γ_ℓ as the *(lower) curvature parameter*, and to the quantity τ_ℓ as the *tolerance parameter*. The set Ω' corresponds to a suitably chosen subset of the space Ω of all possible parameters.[2]

In order to gain intuition for this definition, first suppose that the condition (8) holds with tolerance parameter $\tau_\ell = 0$. In this case, the regularizer plays no role in the definition, and condition (8) is equivalent to the usual definition of strong convexity on the optimization set Ω. As discussed previously, this type of global strong convexity typically *fails* to hold for high-dimensional inference problems. In contrast, when tolerance parameter τ_ℓ is strictly positive, the condition (8) is much milder, in that it only applies to a *limited set* of vectors. For a given pair $\theta \neq \theta'$, consider the inequality

$$\frac{\mathcal{R}^2(\theta-\theta')}{\|\theta-\theta'\|^2} < \frac{\gamma_\ell}{2\,\tau_\ell(\mathcal{L}_n)}. \tag{9}$$

If this inequality is violated, then the right-hand side of the bound (8) is non-positive, in which case the RSC constraint (8) is vacuous. Thus, RSC imposes a non-trivial constraint only on pairs $\theta \neq \theta'$ for which the inequality (9) holds, and a central part of our analysis will be to prove that for our methods, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies a constraint of the form (9). We note that since the regularizer $\mathcal{R}$ is convex, strong convexity of the loss function $\mathcal{L}_n$ also implies the strong convexity of the regularized loss ϕ_n.

We also specify an analogous notion of restricted smoothness:

Definition 2 (**Restricted smoothness (RSM)**). *We say the loss function $\mathcal{L}_n$ satisfies restricted smoothness with respect to $\mathcal{R}$ and with pa-*

[2] As pointed out by a referee, our RSC condition is an instance of the general theory of paraconvexity (e.g., [32]); however, we are not aware of convergence rates for minimizing general paraconvex functions.

rameters $(\gamma_u, \tau_u(\mathcal{L}_n))$ *over the set* Ω' *if*

$$\mathcal{T}_{\mathcal{L}}(\theta;\theta') \leq \frac{\gamma_u}{2}\,\|\theta-\theta'\|^2 + \tau_u(\mathcal{L}_n)\,\mathcal{R}^2(\theta-\theta') \qquad \textit{for all } \theta,\theta' \in \Omega'. \tag{10}$$

As with our definition of restricted strong convexity, the additional tolerance $\tau_u(\mathcal{L}_n)$ is not present in analogous smoothness conditions in the optimization literature, but it is essential in our set-up.

2.3. *Decomposable regularizers.* In past work on the statistical properties of regularization, the notion of a decomposable regularizer has been shown to be useful [27]. Although the focus of this paper is a rather different set of questions—namely, optimization as opposed to statistics—decomposability also plays an important role here. Decomposability is defined with respect to a pair of subspaces defined with respect to the parameter space $\Omega \subseteq \mathbb{R}^d$. The set $\mathcal{M}$ is known as the *model subspace*, whereas the set $\overline{\mathcal{M}}^{\perp}$, referred to as the *perturbation subspace*, captures deviations from the model subspace.

Definition 3. *Given a subspace pair* $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ *such that* $\mathcal{M} \subseteq \overline{\mathcal{M}}$, *we say that a norm* $\mathcal{R}$ *is* $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$-decomposable *if*

$$\mathcal{R}(\alpha+\beta) = \mathcal{R}(\alpha) + \mathcal{R}(\beta) \qquad \textit{for all } \alpha \in \mathcal{M} \textit{ and } \beta \in \overline{\mathcal{M}}^{\perp}. \tag{11}$$

To gain some intuition for this definition, note that by triangle inequality, we always have the bound $\mathcal{R}(\alpha+\beta) \leq \mathcal{R}(\alpha) + \mathcal{R}(\beta)$. For a decomposable regularizer, this inequality always holds with equality. Thus, given a fixed vector $\alpha \in \mathcal{M}$, the key property of any decomposable regularizer is that it affords the *maximum penalization* of any deviation $\beta \in \overline{\mathcal{M}}^{\perp}$.

For a given error norm $\|\cdot\|$, its interaction with the regularizer $\mathcal{R}$ plays an important role in our results. In particular, we have the following:

Definition 4 (Subspace compatibility). *Given the regularizer* $\mathcal{R}(\cdot)$ *and a norm* $\|\cdot\|$, *the associated* subspace compatibility *is given by*

$$\Psi(\overline{\mathcal{M}}) := \sup_{\theta \in \bar{\mathcal{M}}\setminus\{0\}} \frac{\mathcal{R}(\theta)}{\|\theta\|} \qquad \textit{when } \overline{\mathcal{M}} \neq \{0\}, \qquad \textit{and } \Psi(\{0\}) := 0. \tag{12}$$

The quantity $\Psi(\overline{\mathcal{M}})$ corresponds to the Lipschitz constant of the norm $\mathcal{R}$ with respect to $\|\cdot\|$, when restricted to the subspace $\overline{\mathcal{M}}$.

2.4. *Some illustrative examples.* We now describe some particular examples of M-estimators with decomposable regularizers, and discuss the form of the projected gradient updates as well as RSC/RSM conditions. We cover two main families of examples: log-linear models with sparsity constraints and ℓ_1-regularization (Section 2.4.1), and matrix regression problems with nuclear norm regularization (Section 2.4.2).

2.4.1. *Sparse log-linear models and ℓ_1-regularization.* Suppose that each sample Z_i consists of a scalar-vector pair $(y_i, x_i) \in \mathbb{R} \times \mathbb{R}^d$, corresponding to the scalar response $y_i \in \mathcal{Y}$ associated with a vector of predictors $x_i \in \mathbb{R}^d$. A log-linear model with canonical link function assumes that the response y_i is linked to the covariate vector x_i via a conditional distribution of the form $\mathbb{P}(y_i \mid x_i; \theta^*, \sigma) \propto \exp\Big\{\frac{y_i \langle \theta^*, x_i\rangle - \Phi(\langle \theta^*, x_i\rangle)}{c(\sigma)}\Big\}$, where $c(\sigma)$ is a known scaling parameter, $\Phi(\cdot)$ is a known link function, and $\theta^* \in \mathbb{R}^d$ is an unknown regression vector. In many applications, θ^* is relatively sparse, so that it is natural to impose an ℓ_1-constraint. Computing the maximum likelihood estimate subject to such a constraint involves solving the convex program[3]

$$(13) \quad \widehat{\theta} \in \arg\min_{\theta \in \Omega} \underbrace{\Big\{\frac{1}{n}\sum_{i=1}^{n} \big\{\Phi(\langle \theta,\, x_i\rangle) - y_i \langle \theta,\, x_i\rangle\big\}\Big\}}_{\mathcal{L}_n(\theta; Z_1^n)} \quad \text{such that } \|\theta\|_1 \leq \rho,$$

with $x_i \in \mathbb{R}^d$ as its i^{th} row. We refer to this estimator as the log-linear Lasso; it is a special case of the M-estimator (1). Ordinary linear regression is the special case of the log-linear setting with $\Phi(t) = t^2/2$ and $\Omega = \mathbb{R}^d$, and in this case, the estimator (13) corresponds to ordinary least-squares version of Lasso [13, 40]. Other forms of log-linear Lasso that are of interest include logistic regression, Poisson regression, and multinomial regression.

Projected gradient updates:. For the log-linear loss from equation (13), an easy calculation yields the gradient $\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n x_i \big\{\Phi'(\langle \theta,\, x_i\rangle - y_i)\big\}$, and the update (5) corresponds to the Euclidean projection of the vector $\theta^t - \frac{1}{\gamma_u}\nabla \mathcal{L}_n(\theta^t)$ onto the ℓ_1-ball of radius ρ. It is well-known that this projection can be characterized in terms of soft-thresholding, and that the projected update (5) can be computed in $\mathcal{O}(d)$ operations [14].

Composite gradient updates:. The composite gradient update for this problem amounts to solving

$$\theta^{t+1} = \arg\min_{\|\theta\|_1 \leq \bar{\rho}} \Big\{\langle \theta,\, \nabla \mathcal{L}_n(\theta)\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|_2^2 + \lambda_n \|\theta\|_1\Big\}.$$

[3] Φ is convex since it is the log-partition function of a canonical exponential family.

The update can be computed by two soft-thresholding operations. The first step is soft thresolding the vector $\theta^t - \frac{1}{\gamma_u}\nabla\mathcal{L}_n(\theta^t)$ at a level λ_n. If the resulting vector has ℓ_1-norm greater than $\bar{\rho}$, then we project on to the ℓ_1-ball just like before. Overall, the complexity of the update is still $\mathcal{O}(d)$ as before.

Decomposability of ℓ_1-norm:. We now illustrate how the ℓ_1-norm is decomposable with respect to appropriately chosen subspaces. For any subset $S \subseteq \{1, 2, \ldots, d\}$, consider the subspace

$$\mathcal{M}(S) := \big\{\alpha \in \mathbb{R}^d \mid \alpha_j = 0 \quad \text{for all } j \notin S\big\}, \tag{14}$$

corresponding to all vectors supported only on S. Defining $\overline{\mathcal{M}}(S) = \mathcal{M}(S)$, its orthogonal complement (with respect to the Euclidean inner product) is given by $\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S) = \big\{\beta \in \mathbb{R}^d \mid \beta_j = 0 \text{ for all } j \in S\big\}$. Since any pair of vectors $\alpha \in \mathcal{M}(S)$ and $\beta \in \overline{\mathcal{M}}^{\perp}(S)$ have disjoint supports, it follows that $\|\alpha\|_1 + \|\beta\|_1 = \|\alpha + \beta\|_1$. Consequently, for any subset S, the ℓ_1-norm is decomposable with respect to the pairs $(\mathcal{M}(S), \mathcal{M}^{\perp}(S))$.

In analogy to the ℓ_1-norm, various types of group-sparse norms are also decomposable with respect to non-trivial subspace pairs. We refer the reader to the paper [27] for further examples of such decomposable norms.

RSC/RSM conditions:. A calculation using the mean-value theorem shows that for the loss function (13), the error in the first-order Taylor series, as previously defined in equation (6), can be written as

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') = \frac{1}{n}\sum_{i=1}^{n} \Phi''\big(\langle\theta_t,\, x_i\rangle\big)\, \big(\langle x_i,\, \theta - \theta'\rangle\big)^2$$

where $\theta_t = t\theta + (1-t)\theta'$ for some $t \in [0, 1]$. When $n < d$, then we can always find pairs $\theta \neq \theta'$ such that $\langle x_i,\, \theta - \theta'\rangle = 0$ for all $i = 1, 2, \ldots, n$, showing that the objective function can never be strongly convex. On the other hand, RSC for log-linear models requires only that there exist positive numbers $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ such that for all $\theta, \theta' \in \Omega'$

$$\frac{1}{n}\sum_{i=1}^{n} \Phi''\big(\langle\theta_t,\, x_i\rangle\big)\, \big(\langle x_i,\, \theta - \theta'\rangle\big)^2 \geq \frac{\gamma_\ell}{2}\, \|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n)\, \mathcal{R}^2(\theta - \theta'), \tag{15}$$

where $\Omega' := \Omega \cap \mathbb{B}_2(R)$. This restriction is essential because for many generalized linear models (e.g. logistic), the Hessian function Φ'' approaches zero as its argument diverges. RSM imposes an analogous upper bound on the Taylor error. For a broad class of log-linear models, such bounds hold with tolerance $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ of the order $\sqrt{\frac{\log d}{n}}$. A detailed discussion of RSC for exponential families can be found in the paper [27].

In the special case of linear regression, we have $\Phi''(t) = 1$ for all $t \in \mathbb{R}$, so that the lower bound (15) involves only the Gram matrix $X^T X/n$. (Here $X \in \mathbb{R}^{n \times d}$ is the usual design matrix, with $x_i \in \mathbb{R}^d$ as its i^{th} row.) For linear regression and ℓ_1-regularization, the RSC condition is equivalent to

$$(16) \quad \frac{\|X(\theta - \theta')\|_2^2}{n} \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|_2^2 - \tau_\ell(\mathcal{L}_n)\, \|\theta - \theta'\|_1^2 \qquad \text{for all } \theta, \theta' \in \Omega.$$

Such a condition corresponds to a variant of the restricted eigenvalue (RE) conditions that have been studied in the literature [7, 42]. Such RE conditions are significantly milder than the restricted isometry property; we refer the reader to van de Geer and Buhlmann [42] for an in-depth comparison of different RE conditions. From past work, the condition (16) is satisfied with high probability with a constant $\gamma_\ell > 0$ and tolerance $\tau_\ell(\mathcal{L}_n) \asymp \frac{\log d}{n}$ for a broad classes of anisotropic random design matrices [33, 38], and parts of our analysis make use of this fact.

2.4.2. *Matrices and nuclear norm regularization.* We now discuss a general class of matrix regression problems that falls within our framework. Consider the space of $d_1 \times d_2$ matrices endowed with the trace inner product $\langle\!\langle A,\ B \rangle\!\rangle := \text{trace}(A^T B)$. Let $\Theta^* \in \mathbb{R}^{d_1 \times d_2}$ be an unknown matrix and suppose that for $i = 1, 2, \ldots, n$, we observe the pair $Z_i = (y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$, where the scalar response y_i and covariate matrix X_i are linked to the unknown matrix Θ^* via the linear model

$$(17) \qquad y_i = \langle\!\langle X_i,\ \Theta^* \rangle\!\rangle + w_i, \qquad \text{for } i = 1, 2, \ldots, n.$$

Here w_i is an additive observation noise. In many contexts, it is natural to assume that Θ^* is exactly low-rank, or approximately so, meaning that it is well-approximated by a matrix of low rank. In such settings, a number of authors (e.g., [15, 37, 28]) have studied the M-estimator

$$(18) \quad \widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \langle\!\langle X_i,\ \Theta \rangle\!\rangle \big)^2 \right\} \quad \text{such that } |\!|\!|\Theta|\!|\!|_1 \leq \rho,$$

or the corresponding regularized version. Defining $d = \min\{d_1, d_2\}$, the *nuclear or trace norm* is given by $|\!|\!|\Theta|\!|\!|_1 := \sum_{j=1}^{d} \sigma_j(\Theta)$, corresponding to the sum of the singular values. As discussed in Section 3.3, there are various applications in which this estimator and variants thereof have proven useful.

Form of projected gradient descent:. For the M-estimator (18), the projected gradient updates take a very simple form—namely

$$\Theta^{t+1} = \Pi\Big(\Theta^t - \frac{1}{\gamma_u}\frac{\sum_{i=1}^n \big(y_i - \langle\langle X_i,\ \Theta^t\rangle\rangle\big)\, X_i}{n}\Big), \tag{19}$$

where Π denotes Euclidean (i.e. in Frobenius norm) projection onto the nuclear norm ball $\mathbb{B}_N(\rho) = \{\Theta \in \mathbb{R}^{d_1\times d_2} \mid |||\Theta|||_1 \leq \rho\}$. This nuclear norm projection can be obtained by first computing the singular value decomposition (SVD), and then projecting the vector of singular values onto the ℓ_1-ball. The latter step can be achieved by the fast projection algorithms discussed earlier, and there are various methods for fast computation of SVDs. The composite gradient update also has a simple form, requiring at most two singular value thresholding operations.

Decomposability of nuclear norm:. We now define matrix subspaces for which the nuclear norm is decomposable. Defining $d := \min\{d_1, d_2\}$, let $U \in \mathbb{R}^{d_1\times d}$ and $V \in \mathbb{R}^{d_2\times d}$ be arbitrary matrices with orthonormal columns. Using col to denote the column span of a matrix, we define the subspaces[4]

$$\mathcal{M}(U,V) := \big\{\Theta \in \mathbb{R}^{d_1\times d_2} \mid \mathrm{col}(\Theta^T) \subseteq \mathrm{col}(V),\, \mathrm{col}(\Theta) \subseteq \mathrm{col}(U)\big\}, \quad \text{and}$$
$$\overline{\mathcal{M}}^{\perp}(U,V) := \big\{\Theta \in \mathbb{R}^{d_1\times d_2} \mid \mathrm{col}(\Theta^T) \subseteq (\mathrm{col}(V))^{\perp},\, \mathrm{col}(\Theta) \subseteq (\mathrm{col}(U))^{\perp}\big\}.$$

Finally, let us verify the decomposability of the nuclear norm . By construction, any pair of matrices $\Theta \in \mathcal{M}(U,V)$ and $\Gamma \in \overline{\mathcal{M}}^{\perp}(U,V)$ have orthogonal row and column spaces, which implies the required decomposability condition—namely $|||\Theta + \Gamma|||_1 = |||\Theta|||_1 + |||\Gamma|||_1$.

Finally, we note that in some special cases such as matrix completion or matrix decomposition, Ω' will involve an additional bound on the entries of Θ^* as well as the iterates Θ^t to establish RSC/RSM conditions.

3. Main results and some consequences.

We are now equipped to state the two main results of our paper, and discuss some of their consequences. We illustrate its application to several statistical models, including sparse regression (Section 3.2), matrix estimation with rank constraints (Section 3.3), and matrix decomposition problems (Section 3.4). The proofs of all our results can be found in the supplementary material [2].

[4] Note that the model space $\mathcal{M}(U,V)$ is *not equal* to $\overline{\mathcal{M}}(U,V)$. Nonetheless, as required by Definition 3, we do have the inclusion $\mathcal{M}(U,V) \subseteq \overline{\mathcal{M}}(U,V)$.

3.1. *Geometric convergence.* Recall that the projected gradient algorithm (3) is well-suited to solving an M-estimation problem in its constrained form, whereas the composite gradient algorithm (4) is appropriate for a regularized problem. Accordingly, let $\widehat{\theta}$ be any optimum of the constrained problem (1), or the regularized problem (2), and let $\{\theta^t\}_{t=0}^{\infty}$ be a sequence of iterates generated by generated by the projected gradient (3), or the the composite gradient updates (4), respectively. Of primary interest to us are bounds on the *optimization error*, which can be measured either in terms of the error vector $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, or the difference between the objective values at θ^t and $\widehat{\theta}$. In this section, we state two main results—Theorems 1 and 2—corresponding to the constrained and regularized cases respectively. In addition to the optimization error previously discussed, both of these results involve the *statistical error* $\Delta^* := \widehat{\theta} - \theta^*$ between the optimum $\widehat{\theta}$ and the nominal parameter θ^*. At a high level, these results guarantee that under the RSC/RSM conditions, the optimization error shrinks geometrically, with a contraction coefficient that depends on the the loss function $\mathcal{L}_n$ via the parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$. An interesting feature is that the contraction occurs only up to a certain tolerance ϵ^2 depending on these same parameters, and the statistical error. However, as we discuss, for many statistical problems of interest, we can show that this tolerance ϵ^2 is of a lower order than the intrinsic statistical error, and consequently our theory gives an upper bound on the number of iterations required to solve an M-estimation problem up to the statistical precision.

Convergence rates for projected gradient:. We now provide the notation necessary for a precise statement of this claim. Our main result involves a family of upper bounds, one for each pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ of $\mathcal{R}$-decomposable subspaces (see Defn. 3). This subspace choice can be optimized for different model to obtain the tightest possible bounds. For a given pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ such that $16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n) < \gamma_u$, let us define the *contraction coefficient*

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\} \Big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\gamma_u}\Big\}^{-1}. \tag{21}$$

In addition, we define the *tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\ \big(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \Psi(\overline{\mathcal{M}})\|\Delta^*\| + 2\mathcal{R}(\Delta^*)\big)^2}{\gamma_u}, \tag{22}$$

where $\Delta^* = \widehat{\theta} - \theta^*$ is the statistical error, and $\Pi_{\mathcal{M}^\perp}(\theta^*)$ denotes the Euclidean projection of θ^* onto the subspace $\mathcal{M}^\perp$.

In terms of these two ingredients, we now state our first main result:

THEOREM 1. *Suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ and $(\gamma_u, \tau_u(\mathcal{L}_n))$ respectively. Let $(\mathcal{M}, \overline{\mathcal{M}})$ be any $\mathcal{R}$-decomposable pair of subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$ and*

$$0 < \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1. \tag{23}$$

Then for any optimum $\widehat{\theta}$ of the problem (1) *for which the constraint is active, for all iterations $t = 0, 1, 2, \ldots,$ we have*

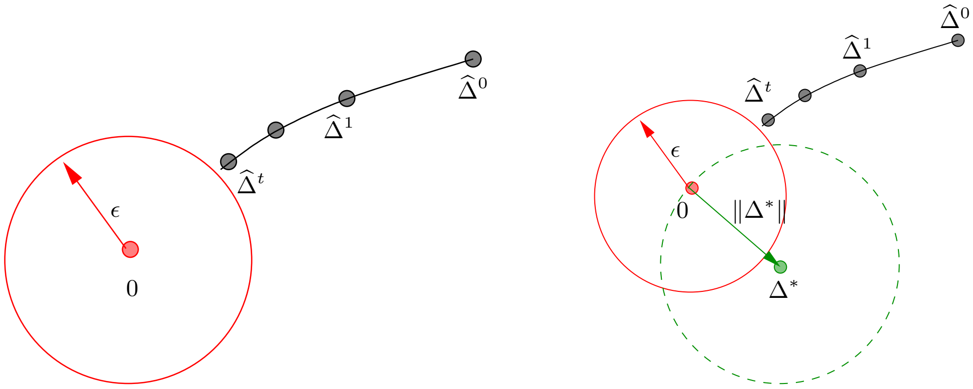
$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{1 - \kappa}, \tag{24}$$

where $\kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}})$.

Remarks:. Theorem 1 actually provides a family of upper bounds, one for each $\mathcal{R}$-decomposable pair $(\mathcal{M}, \overline{\mathcal{M}})$ such that condition (23) holds. This condition is always satisfied by setting $\overline{\mathcal{M}}$ equal to the trivial subspace $\{0\}$: indeed, by definition (12) of the subspace compatibility, we have $\Psi(\overline{\mathcal{M}}) = 0$, and hence $\kappa(\mathcal{L}_n; \{0\}) = \big(1 - \frac{\gamma_\ell}{\gamma_u}\big) < 1$. Although this choice of $\overline{\mathcal{M}}$ minimizes the contraction coefficient, it will lead[5] to a very large tolerance parameter $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$. A more typical application of Theorem 1 involves non-trivial choices of the subspace $\overline{\mathcal{M}}$.

The bound (24) guarantees that the optimization error decreases geometrically, with contraction factor $\kappa \in (0, 1)$, up to a certain tolerance proportional to $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as illustrated in Figure 2(a). Whenever the tolerance terms in the RSC/RSM conditions decay to zero as the sample size increases—the typical case— then the contraction factor κ approaches $1 - \gamma_\ell/\gamma_u$. The appearance of the ratio γ_ℓ/γ_u is natural since it measures the conditioning of the objective function; more specifically, it is essentially a restricted condition number of the Hessian matrix. On the other hand, the residual error ϵ defined in equation (22) depends on the choice of decomposable subspaces, the parameters of the RSC/RSM conditions, and the statistical error $\Delta^* = \widehat{\theta} - \theta^*$. In the corollaries of Theorem 1 to follow, we show that the subspaces can often be chosen such that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) = o(\|\widehat{\theta} - \theta^*\|^2)$. Consequently, the bound (24) guarantees geometric convergence up to a residual error *smaller than statistical precision*, as illustrated in Figure 2(b). This is sensible, since in statistical settings, there is no point to optimizing beyond the statistical precision.

[5] Indeed, the setting $\mathcal{M}^\perp = \mathbb{R}^d$ means that the term $\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) = \mathcal{R}(\theta^*)$ appears in the tolerance; this quantity is far larger than statistical precision.



(a) (b)

Fig 2. (a) Generic illustration of Theorem 1. The optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ is guaranteed to decrease geometrically with coefficient $\kappa \in (0,1)$, up to the tolerance $\epsilon^2 = \epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, represented by the circle. (b) Relation between the optimization tolerance $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ (solid circle) and the statistical precision $\|\Delta^*\| = \|\theta^* - \widehat{\theta}\|$ (dotted circle). In many settings, we have $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \ll \|\Delta^*\|^2$.

The result of Theorem 1 takes a simpler form when there is a subspace $\mathcal{M}$ that includes θ^*, and the $\mathcal{R}$-ball radius is chosen such that $\rho \leq \mathcal{R}(\theta^*)$.

Corollary 1. *In addition to the conditions of Theorem 1, suppose that $\theta^* \in \mathcal{M}$ and $\rho \leq \mathcal{R}(\theta^*)$. Then as long as $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$, we have for all iterations $t = 0, 1, 2, \ldots,$*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + o\big(\|\widehat{\theta} - \theta^*\|^2\big). \tag{25}$$

Thus, Corollary 1 guarantees that the optimization error decreases geometrically, with contraction factor κ, up to a tolerance that is of strictly lower order than the statistical precision $\|\widehat{\theta} - \theta^*\|^2$. As will be clarified in several examples to follow, the condition $\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big) = o(1)$ is satisfied for many statistical models, including sparse linear regression and low-rank matrix regression. This result is illustrated in Figure 2(b), where the solid circle represents the optimization tolerance, and the dotted circle represents the statistical precision. In the results to follow, we quantify the term $o\big(\|\widehat{\theta} - \theta^*\|^2\big)$ in a more precise manner for different statistical models.

Convergence rates for composite gradient:. We now present our main result for the composite gradient iterates (4) that are suitable for the Lagrangian-based estimator (2). As before, our analysis yields a range of bounds indexed by subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ that are $\mathcal{R}$-decomposable. For any subspace $\overline{\mathcal{M}}$ such that $64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}) < \gamma_\ell$, we define *effective RSC coefficient* as

$$\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}). \tag{26}$$

This coefficient accounts for the residual amount of strong convexity after accounting for the lower tolerance terms. In addition, we define the *compound contraction coefficient* as

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \left\{1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\overline{\gamma_\ell}}\right\} \xi(\overline{\mathcal{M}}) \tag{27}$$

where $\xi(\overline{\mathcal{M}}) := \big(1 - \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\big)^{-1}$, and $\Delta^* = \widehat{\theta}_{\lambda_n} - \theta^*$ is the statistical error vector[6] for a specific choice of $\bar{\rho}$ and λ_n. As before, the coefficient κ measures the geometric rate of convergence for the algorithm. Finally, we define the *compound tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := 8\,\xi(\overline{\mathcal{M}})\,\beta(\overline{\mathcal{M}})\left(6\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\right)^2, \tag{28}$$

where $\beta(\overline{\mathcal{M}}) := 2\left(\frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{128\tau_u(\mathcal{L}_n)\Psi^2(\bar{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\tau_\ell(\mathcal{L}_n) + 8\tau_u(\mathcal{L}_n) + 2\tau_\ell(\mathcal{L}_n)$. As with our previous result, the tolerance parameter determines the radius up to which geometric convergence can be attained.

Recall that the regularized problem (2) involves both a regularization weight λ_n, and a constraint radius $\bar{\rho}$. Our theory requires that the constraint radius is chosen such that $\bar{\rho} \geq \mathcal{R}(\theta^*)$, which ensures that θ^* is feasible. In addition, the regularization parameter should be chosen to satisfy

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\theta^*)), \tag{29}$$

where $\mathcal{R}^*$ is the dual norm of the regularizer. This constraint is known to play an important role in proving bounds on the statistical error of regularized M-estimators (see the paper [27] and references therein for further details). Recalling the definition (2) of the overall objective function ϕ_n, the following result provides bounds on the *excess loss* $\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n})$.

THEOREM 2. *Consider the optimization problem* (2) *for a radius $\bar{\rho}$ such that θ^* is feasible, and a regularization parameter λ_n satisfying the bound* (29), *and suppose that the loss function $\mathcal{L}_n$ satisfies the RSC/RSM condition with*

[6] When the context is clear, we remind the reader that we drop the subscript λ_n on the parameter $\widehat{\theta}$.

parameters $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ *and* $(\gamma_u, \tau_u(\mathcal{L}_n))$ *respectively. Let* $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ *be any* $\mathcal{R}$*-decomposable pair such that*

$$\kappa \;\equiv\; \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) \in [0,1), \quad \textit{and} \quad \frac{32\,\bar{\rho}}{1-\kappa(\mathcal{L}_n;\overline{\mathcal{M}})}\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}}) \;\leq\; \lambda_n. \tag{30}$$

Then for any $\delta^2 \geq \frac{\epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})}{(1-\kappa)}$*, we have* $\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n}) \leq \delta^2$ *for all*

$$t \;\geq\; \frac{2\log\frac{\phi_n(\theta^0)-\phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}}{\log(1/\kappa)} + \log_2\log_2\left(\frac{\bar{\rho}\lambda_n}{\delta^2}\right)\left(1+\frac{\log 2}{\log(1/\kappa)}\right). \tag{31}$$

Remarks:. Note that the bound (31) guarantees the excess loss $\phi_n(\theta^t) - \phi_n(\widehat{\theta})$ decays geometrically up to any squared error δ^2 larger than the compound tolerance (28). Moreover, the RSC condition also allows us to further translate this result to a bound on the optimization error $\theta^t - \widehat{\theta}$. In particular, for any iterate θ^t such that $\phi_n(\theta^t) - \phi_n(\widehat{\theta}) \leq \delta^2$, we are guaranteed that

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|^2 \leq \frac{2\delta^2}{\overline{\gamma_\ell}} + \frac{16\delta^2\tau_\ell(\mathcal{L}_n)}{\overline{\gamma_\ell}\lambda_n^2} + \frac{4\tau_\ell(\mathcal{L}_n)(6\Psi(\overline{\mathcal{M}}) + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)))^2}{\overline{\gamma_\ell}}. \tag{32}$$

In conjunction with Theorem 2, we see that it suffices to take a number of steps that is logarithmic in the inverse tolerance $(1/\delta)$, again showing a geometric rate of convergence.

Whereas Theorem 1 requires setting the radius so that the constraint is active, Theorem 2 has only a very mild constraint on the radius $\bar{\rho}$, namely that it be large enough such that $\bar{\rho} \geq \mathcal{R}(\theta^*)$. The reason for this much milder requirement is that the additive regularization with weight λ_n suffices to constrain the solution, whereas the extra side constraint is only needed to ensure good behavior of the optimization algorithm in the first few iterations.

Step-size setting:. It seems that the updates (3) and (4) need to know the smoothness bound γ_u in order to set the step-size for gradient updates. However, we can use the same doubling trick as described in Algorithm (3.1) of Nesterov [31]. At each step, we check if the smoothness upper bound holds at the current iterate relative to the previous one. If the condition does not hold, we double our estimate of γ_u and resume. Nesterov [31] demonstrates that this guarantees a geometric convergence with a contraction factor worse at most by a factor of 2, compared to the knowledge of γ_u.

The following subsections are devoted to the development of some consequences of Theorems 1 and 2 and Corollary 1 for some specific statistical models, among them sparse linear regression with ℓ_1-regularization, and matrix regression with nuclear norm regularization. In contrast to the entirely deterministic arguments that underlie the Theorems 1 and 2, these corollaries involve probabilistic arguments, more specifically in order to establish that the RSC and RSM properties hold with high probability.

3.2. *Sparse vector regression.* Recall from Section 2.4.1 the observation model for sparse linear regression. In a variety of applications, it is natural to assume that θ^* is sparse. For a parameter $q \in [0,1]$ and radius $R_q > 0$, let us define the ℓ_q "ball"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \leq R_q\big\}. \tag{33}$$

Note that $q = 0$ corresponds to the case of "hard sparsity", for which any vector $\beta \in \mathbb{B}_0(R_0)$ is supported on a set of cardinality at most R_0. For $q \in (0,1]$, membership in the set $\mathbb{B}_q(R_q)$ enforces a decay rate on the ordered coefficients, thereby modelling approximate sparsity. In order to estimate the unknown regression vector $\theta^* \in \mathbb{B}_q(R_q)$, we consider the least-squares Lasso estimator from Section 2.4.1, based on $\mathcal{L}(\theta; Z_1^n) := \frac{1}{2n}\|y - X\theta\|_2^2$, where $X \in \mathbb{R}^{n\times d}$ is the design matrix. In order to state a concrete result, we consider a random design matrix X, in which each row $x_i \in \mathbb{R}^d$ is drawn i.i.d. from a $N(0,\Sigma)$ distribution, where Σ is the covariance matrix. We use $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ to refer the maximum and minimum eigenvalues of Σ respectively, and $\zeta(\Sigma) := \max\limits_{j=1,2,\ldots,d} \Sigma_{jj}$ for the maximum variance. We also assume that the observation noise is zero-mean and ν^2-sub-Gaussian.

Guarantees for constrained Lasso:. Our convergence rate on the optimization error $\theta^t - \widehat{\theta}$ is stated in terms of the contraction coefficient

$$\kappa := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\}\Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{34}$$

where we have adopted the shorthand

$$\chi_n(\Sigma) := \begin{cases} \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\, R_q \left(\frac{\log d}{n}\right)^{1-q/2} & \text{for } q > 0 \\ \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)}\, s\left(\frac{\log d}{n}\right) & \text{for } q = 0 \end{cases}, \tag{35}$$

for a numerical constant c_0. We assume that $\chi_n(\Sigma)$ is small enough to ensure that $\kappa \in (0,1)$; in terms of the sample size, this amounts to a condition of the form $n = \Omega(R_q^{1/(1-q/2)} \log d)$. Such a scaling is sensible, since it is known from minimax theory on sparse linear regression [34] to be necessary for any method to be statistically consistent over the ℓ_q-ball.
With this set-up, we have the following consequence of Theorem 1:

COROLLARY 2 (Sparse vector recovery). *Under conditions of Theorem 1, suppose that we solve the constrained Lasso with* $\rho \leq \|\theta^*\|_1$ *and* $\gamma_u = 2\sigma_{\max}(\Sigma)$.

(a) Exact sparsity: *Suppose that* θ^* *is supported on a subset of cardinality* s. *Then the iterates* (3) *satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma)\, \|\widehat{\theta} - \theta^*\|_2^2 \tag{36}$$

for all $t = 0, 1, 2, \ldots$ *with probability at least* $1 - \exp(-c_1 \log d)$.

(b) Weak sparsity: *Suppose that* $\theta^* \in \mathbb{B}_q(R_q)$ *for some* $q \in (0, 1]$. *Then the error* $\|\theta^t - \widehat{\theta}\|_2^2$ *in the iterates* (3) *is at most*

$$\|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma) \left\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\widehat{\theta} - \theta^*\|_2^2 \right\} \tag{37}$$

for all $t = 0, 1, 2, \ldots$ *with probability at least* $1 - \exp(-c_1 \log d)$.

We can now compare part (a), which deals with the special case of exactly sparse vectors, to some past work that has established convergence guarantees for optimization algorithms for sparse linear regression. Certain methods are known to converge at sublinear rates (e.g., [5]), more specifically at the rate $\mathcal{O}(1/t^2)$. The geometric rate of convergence guaranteed by Corollary 2 is exponentially faster. Other work on sparse regression has provided geometric rates of convergence that hold once the iterates are close to the optimum [9, 17], or geometric convergence up to the noise level ν^2 using various methods, including greedy methods [41] and thresholded gradient methods [16]. In contrast, Corollary 2 guarantees geometric convergence for all iterates up to a precision below that of statistical error. For these problems, the statistical error $\frac{\nu^2 s \log d}{n}$ is typically much smaller than the noise variance ν^2, and decreases as the sample size is increased.

In addition, Corollary 2 also applies to the case of approximately sparse vectors, lying within the set $\mathbb{B}_q(R_q)$ for $q \in (0, 1]$. There are some important differences between the case of exact sparsity and that of approximate sparsity. Part (a) guarantees geometric convergence to a tolerance depending only on the statistical error $\|\widehat{\theta} - \theta^*\|_2$. In contrast, the second result also has the additional term $R_q\big(\frac{\log d}{n}\big)^{1-q/2}$. This second term arises due to the statistical non-identifiability of linear regression over the ℓ_q-ball, and it is no larger than $\|\widehat{\theta} - \theta^*\|_2^2$ with high probability. This fact follows from known results [34] about minimax rates for linear regression over ℓ_q-balls; these unimprovable rates include a term of this order.

Guarantees for regularized Lasso:. Using similar methods, we can also use Theorem 2 to obtain an analogous guarantee for the regularized Lasso estimator. Here focus only on the case of exact sparsity, although the result extends to approximate sparsity in a similar fashion. Letting $c_i, i = 0, 1, 2, 3, 4$ be universal positive constants, we define the modified curvature constant $\overline{\gamma_\ell} := \gamma_\ell - c_0 \, \frac{s \, \log d}{n} \zeta(\Sigma)$. Our results assume that $n = \Omega(s \log d)$, a condition known to be necessary for statistical consistency, so that $\overline{\gamma_\ell} > 0$. The contraction factor then takes the form

$$\kappa := \big\{1 - \frac{\sigma_{\min}(\Sigma)}{16 \sigma_{\max}(\Sigma)} + c_1 \chi_n(\Sigma)\big\} \, \big\{1 - c_2 \chi_n(\Sigma)\big\}^{-1},$$

where $\chi_n(\Sigma) := \frac{\zeta(\Sigma)}{\overline{\gamma_\ell}} \, \frac{s \, \log d}{n}$. The residual error in the optimization is given by

$$\epsilon_{\text{tol}}^2 := \frac{5 + c_2 \chi_n(\Sigma)}{1 - c_3 \chi_n(\Sigma)} \, \frac{\zeta(\Sigma) \, s \log d}{n} \|\theta^* - \widehat{\theta}\|_2^2, \tag{38}$$

where $\theta^* \in \mathbb{R}^d$ is the unknown regression vector, and $\widehat{\theta}$ is any optimal solution. With this notation, we have the following corollary.

COROLLARY 3 (Regularized Lasso). *Under the conditions of Theorem 2, suppose that we solve the regularized Lasso with $\lambda_n = 6\nu \sqrt{\frac{\log d}{n}}$, and that θ^* is supported on a subset of cardinality at most s. Suppose further that we have*

$$64 \bar{\rho} \frac{\log d}{n} \left\{ 5 + \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64 s \log d / n}{\overline{\gamma_\ell}} \right\} \left\{ \frac{\overline{\gamma_\ell}}{4\gamma_u} - \frac{128 s \log d / n}{\overline{\gamma_\ell}} \right\}^{-1} \leq \lambda_n. \tag{39}$$

Then for any $\delta^2 \geq \epsilon_{tol}^2$ and any optimum $\widehat{\theta}_{\lambda_n}$, we have

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|_2^2 \leq \delta^2 \qquad \textit{for all iterations } t \;\geq\; \big(\log \tfrac{\phi_n(\theta^0) - \phi_n(\widehat{\theta}_{\lambda_n})}{\delta^2}\big) / \big(\log \tfrac{1}{\kappa}\big)$$

with probability at least $1 - \exp(-c_4 \log d)$.

As with Corollary 2(a), this result guarantees that $\mathcal{O}(\log(1/\epsilon_{\text{tol}}^2))$ iterations are sufficient to obtain an iterate θ^t that is within squared error $\mathcal{O}(\epsilon_{\text{tol}}^2)$ of any optimum $\widehat{\theta}_{\lambda_n}$. The condition (39) is the specialization of equation (30) to the sparse linear regression problem, and imposes an upper bound on admissible settings of $\bar{\rho}$ for our theory. Moreover, whenever $\frac{s \log d}{n} = o(1)$—a condition that is required for statistical consistency of *any method* by known minimax results [34]—the residual error ϵ_{tol}^2 is of lower order than the statistical error $\|\theta^* - \theta\|_2^2$.

3.3. *Matrix regression with rank constraints.* We now turn to estimation of matrices under various types of "soft" rank constraints. Recall the model of matrix regression from Section 2.4.2, and the M-estimator based on least-squares regularized with the nuclear norm (18). So as to reduce notational overhead, here we specialize to square matrices $\Theta^* \in \mathbb{R}^{d \times d}$, so that our observations are of the form

$$y_i = \langle\langle X_i,\ \Theta^* \rangle\rangle + w_i, \quad \text{for } i = 1, 2, \ldots, n, \tag{40}$$

where $X_i \in \mathbb{R}^{d \times d}$ is a matrix of covariates, and $w_i \sim N(0, \nu^2)$ is Gaussian noise. As discussed in Section 2.4.2, the nuclear norm $\mathcal{R}(\Theta) = |||\Theta|||_1 = \sum_{j=1}^d \sigma_j(\Theta)$ is decomposable with respect to appropriately chosen matrix subspaces, and we exploit this fact heavily in our analysis.

We model the behavior of both exactly and approximately low-rank matrices by enforcing a sparsity condition on the vector of singular values. In particular, for a parameter $q \in [0, 1]$, we define the ℓ_q-"ball" of matrices

$$\mathbb{B}_q(R_q) := \bigl\{\Theta \in \mathbb{R}^{d \times d} \mid \sum_{j=1}^d |\sigma_j(\Theta)|^q \leq R_q\bigr\}, \tag{41}$$

where $\sigma_j(\Theta)$ denotes the j^{th} singular value of Θ. Note that if $q = 0$, then $\mathbb{B}_0(R_0)$ consists of the set of all matrices with rank at most $r = R_0$. On the other hand, for $q \in (0, 1]$, the set $\mathbb{B}_q(R_q)$ contains matrices of all ranks, but enforces a relatively fast rate of decay on the singular values.

3.3.1. *Bounds for matrix compressed sensing.* We begin by considering the compressed sensing version of matrix regression, a model first introduced by Recht et al. [36], and later studied by other authors (e.g., [22, 28]). In this model, the observation matrices $X_i \in \mathbb{R}^{d \times d}$ are dense and drawn from some random ensemble. The simplest example is the standard Gaussian ensemble, in which each entry of X_i is drawn i.i.d. as standard normal $N(0, 1)$. Note that X_i is a dense matrix in general; this in an important contrast with the matrix completion setting to follow shortly.

Here we consider a more general ensemble of random matrices X_i, in which each matrix $X_i \in \mathbb{R}^{d \times d}$ is drawn i.i.d. from a zero-mean normal distribution in $\mathbb{R}^{d^2}$ with covariance matrix $\Sigma \in \mathbb{R}^{d^2 \times d^2}$. The setting $\Sigma = I_{d^2 \times d^2}$ recovers the standard Gaussian ensemble studied in past work. As usual, we let $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ define the maximum and minimum eigenvalues of Σ, and we define $\zeta_{\text{mat}}(\Sigma) = \sup_{\|u\|_2 = 1} \sup_{\|v\|_2 = 1} \operatorname{var}\bigl(\langle\langle X,\ uv^T \rangle\rangle\bigr)$, corresponding to the maximal variance of X when projected onto rank one matrices. For the identity ensemble, we have $\zeta_{\text{mat}}(I) = 1$.

We now state a result on the convergence of the updates (19) when applied to a statistical problem involving a matrix $\Theta^* \in \mathbb{B}_q(R_q)$. The convergence rate depends on the contraction coefficient

$$\kappa := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\} \Big\{1 - \chi_n(\Sigma)\Big\}^{-1},$$

where $\chi_n(\Sigma) := \frac{c_1 \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \, R_q\big(\frac{d}{n}\big)^{1-q/2}$ for some universal constant c_1. In the case $q = 0$, corresponding to matrices with rank at most r, note that we have $R_0 = r$. With this notation, we have the following convergence guarantee:

COROLLARY 4 (Low-rank matrix recovery). *Under the conditions of Theorem 1, consider the semidefinite program* (18) *with* $\rho \leq |||\Theta^*|||_1$, *and suppose that we apply the projected gradient updates* (19) *with* $\gamma_u = 2\sigma_{\max}(\Sigma)$.

(a) Exactly low-rank: *Suppose that* Θ^* *has rank* $r < d$. *Then the iterates* (19) *satisfy the bound*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \, \chi_n(\Sigma) \, |||\widehat{\Theta} - \Theta^*|||_F^2 \tag{42}$$

for all $t = 0, 1, 2, \ldots$ *with probability at least* $1 - \exp(-c_0 d)$.

(b) Approximately low-rank: *Suppose that* $\Theta^* \in \mathbb{B}_q(R_q)$ *for some* $q \in (0, 1]$. *Then the iterates* (19) *satisfy*

$$|||\Theta^t - \widehat{\Theta}|||_F^2 \leq \kappa^t \, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \chi_n(\Sigma) \left\{ R_q \left(\frac{d}{n}\right)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2 \right\},$$

for all $t = 0, 1, 2, \ldots$ *with probability at least* $1 - \exp(-c_0 d)$.

Although quantitative aspects of the rates are different, Corollary 4 is analogous to Corollary 2. For the case of exactly low rank matrices (part (a)), geometric convergence is guaranteed up to a tolerance involving the statistical error $|||\widehat{\Theta} - \Theta^*|||_F^2$. For the case of approximately low rank matrices (part (b)), the tolerance term involves an additional factor of $R_q\big(\frac{d}{n}\big)^{1-q/2}$. Again, from known results on minimax rates for matrix estimation [37], this term is known to be of comparable or lower order than the quantity $|||\widehat{\Theta} - \Theta^*|||_F^2$. As before, it is also possible to derive an analogous corollary of Theorem 2 for estimating low-rank matrices; in the interests of space, we leave such a development to the reader.

3.3.2. *Bounds for matrix completion.* In this model, observation y_i is a noisy version of a randomly selected entry $\Theta^*_{a(i),b(i)}$ of the unknown matrix

Θ^*. Applications of this matrix completion problem include collaborative filtering [39], where the rows of the matrix Θ^* correspond to users, and the columns correspond to items (e.g., movies in the Netflix database), and the entry Θ^*_{ab} corresponds to user's a rating of item b. Given observations of only a subset of the entries of Θ^*, the goal is to fill in, or complete the matrix, thereby making recommendations of movies that a user has not yet seen.

Matrix completion can be viewed as a particular case of the matrix regression model (17), in particular by setting $X_i = E_{a(i)b(i)}$, corresponding to the matrix with a single one in position $(a(i), b(i))$, and zeroes in all other positions. Note that these observation matrices are extremely sparse, in contrast to the compressed sensing model. Nuclear-norm based estimators for matrix completion are known to have good statistical properties (e.g., [11, 35, 39, 29]). Here we consider the M-estimator

$$\widehat{\Theta} \in \arg\min_{\Theta \in \Omega} \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \Theta_{a(i)b(i)}\big)^2 \quad \text{such that } |||\Theta|||_1 \leq \rho, \tag{43}$$

where $\Omega = \{\Theta \in \mathbb{R}^{d \times d} \mid \|\Theta\|_\infty \leq \frac{\alpha}{d}\}$ is the set of matrices with bounded elementwise ℓ_∞ norm. This constraint eliminates matrices that are overly "spiky" (i.e., concentrate too much of their mass in a single position); as discussed in the paper [29], such spikiness control is necessary in order to bound the non-identifiable component of the matrix completion model.

Corollary 5 (Matrix completion). *Under the conditions of Theorem 1, suppose that $\Theta^* \in \mathbb{B}_q(R_q)$, and that we solve the program (43) with $\rho \leq |||\Theta^*|||_1$. As long as $n > c_0 R_q^{1/(1-q/2)}\, d \log d$ for a sufficiently large constant c_0, then there is a contraction coefficient $\bar{\kappa}_t \in (0,1)$ that decreases with t such that*

$$|||\Theta^{t+1} - \widehat{\Theta}|||_F^2 \leq \bar{\kappa}_t^t\, |||\Theta^0 - \widehat{\Theta}|||_F^2 + c_2 \Big\{ R_q \big(\frac{\alpha^2 d \log d}{n}\big)^{1-q/2} + |||\widehat{\Theta} - \Theta^*|||_F^2 \Big\} \tag{44}$$

for all iterations $t = 0, 1, 2, \ldots$, with probability at least $1 - \exp(-c_1 d \log d)$.

As with our previous results, the residual optimization error in this result is of the same order as known statistical minimax rates for the matrix completion problem under the soft-rank model described here (cf. Theorem 3 in Negahban and Wainwright [29]). In some cases, the bound on $\|\Theta\|_\infty$ in the algorithm (43) might be unknown, or undesirable. While this constraint is necessary in general [29], it can be avoided if more information such as the sampling distribution (that is, the distribution of X_i) is known and used to construct the estimator. In this case, Koltchinskii et al. [21] use an alternative nuclear-norm penalized estimator for which it is not necessary to directly impose an ℓ_∞ bound on $\widehat{\Theta}$.

Again a similar corollary of Theorem 2 can be derived by combining the proof of Corollary 5 with that of Theorem 2. An interesting aspect of this problem is that the condition 30(b) takes the form $\lambda_n > \frac{c\alpha\sqrt{d \log d/n}}{1-\kappa}$, where α is a bound on $\|\Theta\|_\infty$. This condition is independent of $\bar{\rho}$, and hence, given a sample size as stated in the corollary, the algorithm always converges geometrically for any radius $\bar{\rho} \geq |\!|\!|\Theta^*|\!|\!|_1$.

3.4. *Matrix decomposition problems.* In recent years, various researchers have studied methods for solving the problem of matrix decomposition (e.g., [12, 10, 43, 1, 18]). The basic problem has the following form: given a pair of unknown matrices Θ^* and Γ^*, both lying in $\mathbb{R}^{d_1 \times d_2}$, suppose that we observe a third matrix specified by the model $Y = \Theta^* + \Gamma^* + W$, where $W \in \mathbb{R}^{d_1 \times d_2}$ represents observation noise. Typically the matrix Θ^* is assumed to be low-rank, and some low-dimensional structural constraint is assumed on the matrix Γ^*. For example, the papers [12, 10, 18] consider the setting in which Γ^* is sparse, while Xu et al. [43] consider a column-sparse model, in which only a few of the columns of Γ^* have non-zero entries. In order to illustrate the application of our general result to this setting, here we consider the low-rank plus column-sparse framework [43]. (We note that since the ℓ_1-norm is decomposable, similar results can easily be derived for the low-rank plus entrywise-sparse setting as well.)

Since Θ^* is assumed to be low-rank, as before we use the nuclear norm $|\!|\!|\Theta|\!|\!|_1$ as a regularizer (see Section 2.4.2). We assume that the unknown matrix $\Gamma^* \in \mathbb{R}^{d_1 \times d_2}$ is column-sparse, say with at most $s < d_2$ non-zero columns. A suitable convex regularizer for this matrix structure is based on the *columnwise* $(1,2)$-*norm*, given by

$$\|\Gamma\|_{1,2} := \sum_{j=1}^{d_2} \|\Gamma_j\|_2, \tag{45}$$

where $\Gamma_j \in \mathbb{R}^{d_1}$ denotes the j^{th} column of Γ. Note also that the dual norm is given by the *elementwise* $(\infty,2)$-*norm* $\|\Gamma\|_{\infty,2} = \max_{j=1,\ldots,d_2} \|\Gamma_j\|_2$, corresponding to the maximum ℓ_2-norm over columns.

In order to estimate the unknown pair (Θ^*, Γ^*), we consider the M-estimator $(\widehat{\Theta}, \widehat{\Gamma})$ which minimizes the objective

$$\min_{\Theta,\Gamma} |\!|\!|Y - \Theta - \Gamma|\!|\!|_F^2 \text{ s.t. } |\!|\!|\Theta|\!|\!|_1 \leq \rho_\Theta, \;\; \|\Gamma\|_{1,2} \leq \rho_\Gamma, \;\; \|\Theta\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}. \tag{46}$$

The first two constraints restrict Θ and Γ to a nuclear norm ball of radius ρ_Θ and a $(1,2)$-norm ball of radius ρ_Γ, respectively. The final constraint controls the "spikiness" of the low-rank component Θ, as measured in the $(\infty, 2)$-norm, corresponding to the maximum ℓ_2-norm over the columns. As with the elementwise ℓ_∞-bound for matrix completion, this additional constraint is required in order to limit the non-identifiability in matrix decomposition. (See the paper [1] for more discussion of non-identifiability issues in matrix decomposition.)

With this set-up, consider the projected gradient algorithm when applied to the matrix decomposition problem: it generates a sequence of matrix pairs (Θ^t, Γ^t) for $t = 0, 1, 2, \ldots$, and the optimization error is characterized in terms of the matrices $\widehat{\Delta}^t_\Theta := \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}^t_\Gamma := \Gamma^t - \widehat{\Gamma}$. Finally, we measure the optimization error at time t in terms of the squared Frobenius error $e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) := |||\widehat{\Delta}^t_\Theta|||_F^2 + |||\widehat{\Delta}^t_\Gamma|||_F^2$, summed across both the low-rank and column-sparse components.

COROLLARY 6 (Matrix decomposition). *Under the conditions of Theorem 1, suppose that* $\|\Theta^*\|_{\infty,2} \leq \frac{\alpha}{\sqrt{d_2}}$ *and* Γ^* *has at most* s *non-zero columns. If we solve the convex program* (46) *with* $\rho_\Theta \leq |||\Theta^*|||_1$ *and* $\rho_\Gamma \leq \|\Gamma^*\|_{1,2}$, *then for all iterations* $t = 0, 1, 2, \ldots$,

$$e^2(\widehat{\Delta}^t_\Theta, \widehat{\Delta}^t_\Gamma) \leq \left(\frac{3}{4}\right)^t e^2(\widehat{\Delta}^0_\Theta, \widehat{\Delta}^0_\Gamma) + c\left(|||\widehat{\Gamma} - \Gamma^*|||_F^2 + \alpha^2 \frac{s}{d_2}\right).$$

This corollary has some unusual aspects, relative to the previous corollaries. First of all, in contrast to the previous results, the guarantee is a deterministic one (as opposed to holding with high probability). More specifically, the RSC/RSM conditions hold deterministic sense, which should be contrasted with the high probability statements given in Corollaries 2-5. Consequently, the effective conditioning of the problem does not depend on sample size and we are guaranteed geometric convergence at a fixed rate, independent of sample size. The additional tolerance term is completely independent of Θ^* and only depends on the column-sparsity of Γ^*.

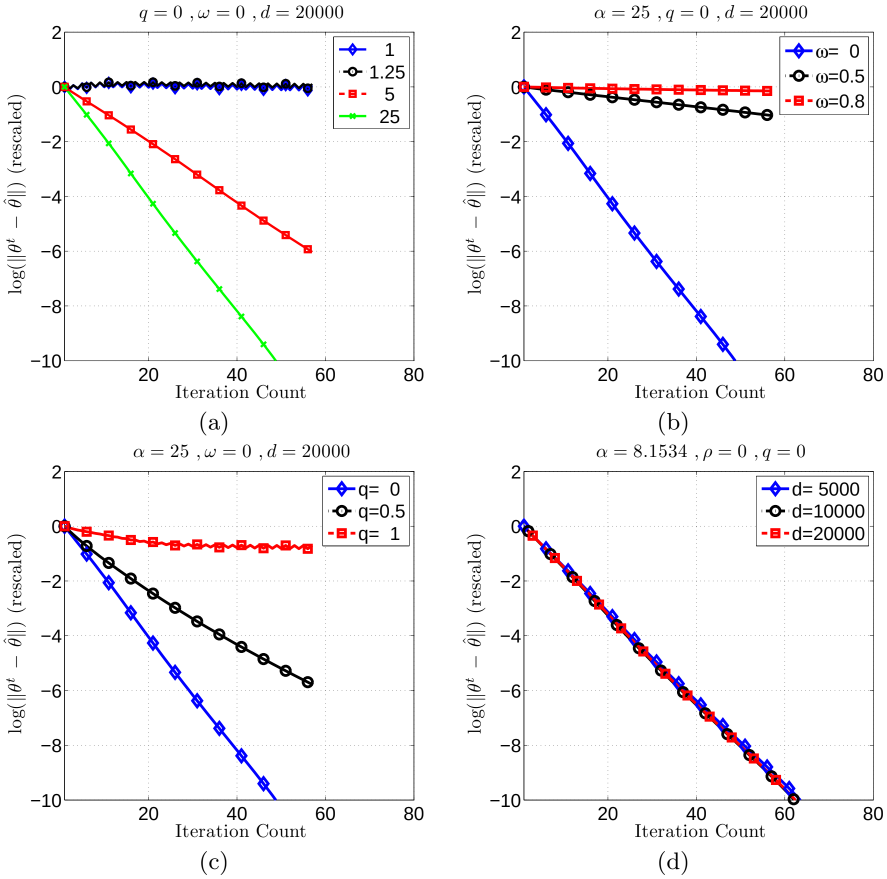
4. Simulation results. In this section, we provide some experimental results that confirm the accuracy of our theoretical results, in particular showing excellent agreement with the linear rates predicted by our theory. In addition, the rates of convergence slow down for smaller sample sizes, which lead to problems with relatively poor conditioning. In all the simulations reported below, we plot the log error $\|\theta^t - \widehat{\theta}\|$ between the iterate θ^t at time t versus the final solution $\widehat{\theta}$. Each curve provides the results averaged over five random trials, according to the ensembles which we now describe.

4.1. *Sparse regression.* We investigate the standard linear regression model $y = X\theta^* + w$ where θ^* is the unknown regression vector belonging to the set $\mathbb{B}_q(R_q)$, and i.i.d. observation noise $w_i \sim N(0, 0.25)$. We consider a family of ensembles for the random design matrix $X \in \mathbb{R}^{n \times d}$. In particular, we construct X by generating each row $x_i \in \mathbb{R}^d$ independently according to following procedure. Let $z_1, \ldots, z_n$ be an i.i.d. sequence of $N(0,1)$ variables, and fix some correlation parameter $\omega \in [0, 1)$. We first initialize by setting $x_{i,1} = z_1/\sqrt{1-\omega^2}$, and then generate the remaining entries by applying the recursive update $x_{i,t+1} = \omega x_{i,t} + z_t$ for $t = 1, 2, \ldots, d-1$, so that $x_i \in \mathbb{R}^d$ is a zero-mean Gaussian random vector. It can be verified that all the eigenvalues of $\Sigma = \mathrm{cov}(x_i)$ lie within the interval $[\frac{1}{(1+\omega)^2}, \frac{2}{(1-\omega)^2(1+\omega)}]$, so that Σ has a a finite condition number for all $\omega \in [0, 1)$. At one extreme, for $\omega = 0$, the matrix Σ is the identity, and so has condition number equal to 1. As $\omega \to 1$, the matrix Σ becomes progressively more ill-conditioned, with a condition number that is very large for ω close to one. As a consequence, although incoherence conditions like the restricted isometry property can be satisfied when $\omega = 0$, they will fail to be satisfied (w.h.p.) once ω is large enough.

For this random ensemble of problems, we have investigated convergence rates for a wide range of dimensions d and radii R_q. Since the results are relatively uniform across the choice of these parameters, here we report results for dimension $d = 20,000$, and radius $R_q = \lceil (\log d)^2 \rceil$. In the case $q = 0$, the radius $R_0 = s$ corresponds to the sparsity level. The per iteration cost in this case is $\mathcal{O}(nd)$. In order to reveal dependence of convergence rates on sample size, we study a range of the form $n = \lceil \alpha \, s \log d \rceil$, where the *order parameter* $\alpha > 0$ is varied.

Our first experiment is based on taking the correlation parameter $\omega = 0$, and the ℓ_q-ball parameter $q = 0$, corresponding to exact sparsity. We then measure convergence rates for sample sizes specified by $\alpha \in \{1, 1.25, 5, 25\}$. As shown by the results plotted in panel (a) of Figure 3, projected gradient descent fails to converge for $\alpha = 1$ or $\alpha = 1.25$; in both these cases, the sample size n is too small for the RSC and RSM conditions to hold, so that a constant step size leads to oscillatory behavior in the algorithm. In contrast, once the order parameter α becomes large enough to ensure that the RSC/RSM conditions hold (w.h.p.), we observe a geometric convergence of the error $\|\theta^t - \widehat{\theta}\|_2$. Moreover the convergence rate is faster for $\alpha = 25$ compared to $\alpha = 5$, since the RSC/RSM constants are better with larger sample size. Such behavior is in agreement with the conclusions of Corollary 2, which predicts that the the convergence rate should improve as the number of samples n is increased.

On the other hand, Corollary 2 also predicts that convergence rates should



(a) (b) (c) (d)

Fig 3. Plot of the log of the optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ in the sparse linear regression problem, rescaled so the plots start at 0. In this problem, $d = 20000$, $s = \lceil \log d \rceil$, $n = \alpha s \log d$. Plot (a) shows convergence for the exact sparse case with $q = 0$ and $\Sigma = I$ (i.e. $\omega = 0$). In panel (b), we observe how convergence rates change as the correlation parameter ω is varied for $q = 0$ and $\alpha = 25$. Plot (c) shows the convergence rates when $\omega = 0$, $\alpha = 25$ and q is varied. Plot (d), repeated from Figure 1 shows that keeping α fixed keeps the converegence rate constant across problem sizes.

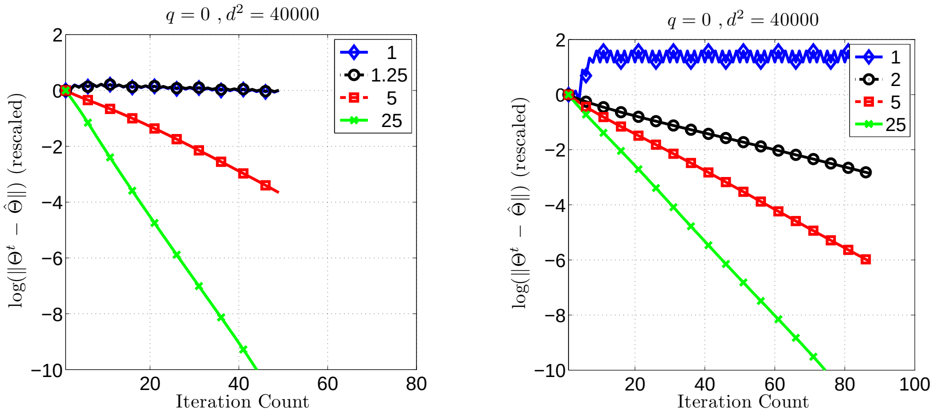
be slower when the condition number of Σ is worse. In order to test this prediction, we again studied an exactly sparse problem ($q = 0$), this time with the fixed sample size $n = \lceil 25 s \log d \rceil$, and we varied the correlation parameter $\omega \in \{0, 0.5, 0.8\}$. As shown in panel (b) of Figure 3, the convergence rates slow down as the correlation parameter is increased and for the case of extremely high correlation of $\omega = 0.8$, the optimization error curve is almost flat—the method makes very slow progress in this case.

A third prediction of Corollary 2 is that the convergence of projected gradient descent should become slower as the sparsity parameter q is varied between exact sparsity ($q = 0$), and the least sparse case ($q = 1$). (In particular, note for $n > \log d$, the quantity χ_n from equation (35) is monotonically increasing with q.) Panel (c) of Figure 3 shows convergence rates for the fixed sample size $n = 25s \log d$ and correlation parameter $\omega = 0$, and with the sparsity parameter $q \in \{0, 0.5, 1.0\}$. As expected, the convergence rate slows down as q increases from 0 to 1. Corollary 2 further captures how the contraction factor changes as the problem parameters (s, d, n) are varied. In particular, it predicts that as we change the triplet simultaneously, while holding the ratio $\alpha = s \log d/n$ constant, the convergence rate should stay the same. This phenomenon that we earlier pointed out in the introduction is indeed demonstrated in Figure 3(d).

5. Low-rank matrix estimation. We also performed experiments with two different versions of low-rank matrix regression. Our simulations applied to instances of the observation model $y_i = \langle\!\langle X_i,\ \Theta^* \rangle\!\rangle + w_i$, for $i = 1, 2, \ldots, n$, where $\Theta^* \in \mathbb{R}^{200 \times 200}$ is a fixed unknown matrix, $X_i \in \mathbb{R}^{200 \times 200}$ is a matrix of covariates, and $w_i \sim N(0, 0.25)$ is observation noise. In analogy to the sparse vector problem, we performed simulations with the matrix Θ^* belonging to the set $\mathbb{B}_q(R_q)$ of approximately low-rank matrices, as previously defined in equation (41) for $q \in [0, 1]$. The case $q = 0$ corresponds to the set of matrices with rank at most $r = R_0$, whereas the case $q = 1$ corresponds to the ball of matrices with nuclear norm at most R_1.

In our first set of matrix experiments, we considered the matrix version of compressed sensing [35], in which each matrix $X_i \in \mathbb{R}^{200 \times 200}$ is randomly formed with i.i.d. $N(0, 1)$ entries, as described in Section 3.3.1. In the case $q = 0$, we formed a matrix $\Theta^* \in \mathbb{R}^{200 \times 200}$ with rank $R_0 = 5$, and performed simulations over the sample sizes $n = \alpha R_0\, d$, with the parameter $\alpha \in \{1, 1.25, 5, 25\}$. The per iteration cost in this case is $\mathcal{O}(nd^2)$. As seen in panel (a) of Figure 4, the projected gradient descent method exhibits behavior that is qualitatively similar to that for the sparse linear regression problem. More specifically, it fails to converge when the sample size (as reflected by the order parameter α) is too small, and converges geometrically with a progressively faster rate as α is increased. We have also observed similar types of scaling as we vary $q \in [0, 1]$.

In our second set of matrix experiments, we studied the behavior of projected gradient descent for the problem of matrix completion, as described in Section 3.3.2. For this problem, we again studied matrices of dimension $d = 200$ and rank $R_0 = 5$, and we varied the sample size as $n = \alpha\, R_0\, d \log d$



(a) (b)

Fig 4. (a) Plot of log Frobenius error $\log(|||\Theta^t - \widehat{\Theta}|||_F)$ versus number of iterations in matrix compressed sensing for a matrix size $d = 200$ with rank $R_0 = 5$, and sample sizes $n = \alpha R_0 d$. For $\alpha \in \{1, 1.25\}$, the algorithm oscillates, whereas geometric convergence is obtained for $\alpha \in \{5, 25\}$. (b) Convergence rate for the matrix completion problem with $d = 200$, $R_0 = 5$, and $n = \alpha R_o d \log(d)$ with $\alpha \in \{1, 2, 5, 25\}$. For $\alpha \in \{2, 5, 25\}$ the algorithm enjoys geometric convergence.

for $\alpha \in \{1, 2, 5, 25\}$. As shown in Figure 4(b), projected gradient descent for matrix completion also enjoys geometric convergence for α large enough.

6. Discussion. In this paper, we have shown that even though high-dimensional M-estimators in statistics are neither strongly convex nor smooth, simple first-order methods can still enjoy global guarantees of geometric convergence. The key insight is that strong convexity and smoothness need only hold in restricted senses, and moreover, these conditions are satisfied with high probability for many statistical models and decomposable regularizers used in practice. Examples include sparse linear regression and ℓ_1-regularization, various statistical models with group-sparse regularization, matrix regression with nuclear norm constraints (including matrix completion and multi-task learning), and matrix decomposition problems. Some related work also shows that related ideas can be used to provide rigorous guarantees for gradient methods in application to certain classes of non-convex programs [23]. Overall, our results highlight some important connections between computation and statistics: the properties of M-estimators favorable for fast rates in statistics can also be used to establish fast rates for optimization algorithms.

Acknowledgements. All three authors were partially supported by grants AFOSR-09NL184; in addition, AA was partially supported by a Microsoft Graduate Fellowship and Google PhD Fellowship, and SN and MJW acknowledge funding from NSF-CDI-0941742. We would like to thank the anonymous reviewers and associate editor for their helpful comments that helped to improve the paper, and Bin Yu for inspiring discussions on the interaction between statistical and optimization error.

REFERENCES

[1] A. Agarwal, S. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Annals of Statistics*, 40(2):1171–1197, 2012.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Supplemental file to Fast global convergence of gradient methods for high-dimensional statistical recovery. Technical report, University of California, Berkeley, 2012.

[3] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 37:2877–2921, 2009.

[4] A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

[5] S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011.

[6] D.P. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 1995.

[7] P. J. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4):1705–1732, 2009.

[8] S. Boyd and L. Vandenberghe. *Convex optimization.* Cambridge University Press, Cambridge, UK, 2004.

[9] K. Bredies and D. A. Lorenz. Linear convergence of iterative soft-thresholding. *Journal of Fourier Analysis and Applications*, 14:813–837, 2008.

[10] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? *J. ACM*, 58:11:1–11:37, 2011.

[11] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Found. Comput. Math.*, 9(6):717–772, 2009.

[12] V. Chandrasekaran, S. Sanghavi, P. Parrilo, and A. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM J. on Optimization*, 21(2):572–596, 2011.

[13] S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Computing*, 20(1):33–61, 1998.

[14] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra. Efficient projections onto the ℓ_1-ball for learning in high dimensions. In *ICML*, 2008.

[15] M. Fazel. *Matrix Rank Minimization with Applications.* PhD thesis, Stanford, 2002. Available online: http://faculty.washington.edu/mfazel/thesis-final.pdf.

[16] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, 2009.

[17] E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for ℓ_1-minimization: Methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

[18] D. Hsu, S.M. Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. *IEEE Trans. Info. Theory*, 57(11):7221 –7234, 2011.

[19] J. Huang and T. Zhang. The benefit of group sparsity. *The Annals of Statistics*, 38(4):1978–2004, 2010.

[20] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

[21] V. Koltchinskii, K. Lounici, and A. B. Tsybakov. Nuclear-norm penalization and optimal rates for noisy low-rank matrix completion. *Annals of Statistics*, 39:2302–2329, 2011.

[22] K. Lee and Y. Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint. Technical report, UIUC, 2009. Available at arXiv:0903.4742.

[23] P. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with non-convexity. *Annals of Statistics*, 2012. To appear; originally posted as http://arxiv.org/abs/1109.3714.

[24] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. In *COLT*, 2009.

[25] Z. Q. Luo and P. Tseng. Error bounds and convergence analysis of feasible descent methods: a general approach. *Annals of Operations Research*, 46-47:157–178, 1993.

[26] N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the Lasso. *Annals of Statistics*, 34:1436–1462, 2006.

[27] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. In *NIPS*, 2009. To appear in Statistical Science.

[28] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, 39(2):1069–1097, 2011.

[29] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13:1665–1697, May 2012.

[30] Y. Nesterov. *Introductory Lectures on Convex Optimization.* Kluwer Academic Publishers, New York, 2004.

[31] Y. Nesterov. Gradient methods for minimizing composite objective function. Technical Report 76, Center for Operations Research and Econometrics (CORE), Catholic University of Louvain (UCL), 2007.

[32] H. V. Ngai and J. P. Penot. Paraconvex functions and paraconvex sets. *Studia Mathematica*, 184:1–29, 2008.

[33] G. Raskutti, M. J. Wainwright, and B. Yu. Restricted eigenvalue conditions for correlated Gaussian designs. *Journal of Machine Learning Research*, 11:2241–2259, August 2010.

[34] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Info. Theory*, 57(10):6976—6994, 2011.

[35] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

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

[37] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. *Annals of Statistics*, 39(2):887–930, 2011.

[38] M. Rudelson and S. Zhou. Reconstruction from anisotropic random measurements. Technical report, University of Michigan, July 2011.

[39] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *NIPS*, 2005.

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

[41] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Info. Theory*, 53(12):4655–4666, 2007.

[42] S. van de Geer and P. Buhlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[43] H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. *IEEE Trans. Info. Theory*, 58(5):3047 –3064, May 2012.

[44] C. H. Zhang and J. Huang. The sparsity and bias of the lasso selection in high-dimensional linear regression. *Annals of Statistics*, 36(4):1567–1594, 2008.

[45] P. Zhao, G. Rocha, and B. Yu. Grouped and hierarchical model selection through composite absolute penalties. *Annals of Statistics*, 37(6A):3468–3497, 2009.

Alekh Agarwal
Department of EECS
University of California Berkeley
Berkeley CA 94720
E-mail: alekh@eecs.berkeley.edu

Sahand Negahban
Department of EECS
Massachusetts Institute of Technology
32 Vassar Street
Cambirdge MA 02139
E-mail: sahandn@mit.edu

Martin J. Wainwright
Department of EECS and Statistics
University of California Berkeley
Berkeley CA 94720
E-mail: wainwrig@stat.berkeley.edu