Fast global convergence of gradient methods for high-dimensional statistical recovery

Alekh Agarwal[1] **Sahand N. Negahban**[1] **Martin J. Wainwright**[1,2]
Department of Electrical Engineering and Computer Science[1] and Department of Statistics[2]
University of California, Berkeley
Berkeley, CA 94720-1776
{alekh,sahand_n,wainwrig}@eecs.berkeley.edu

Abstract

Many statistical M-estimators are based on convex optimization problems formed by the weighted sum of a loss function with a norm-based regularizer. We analyze the convergence rates of first-order gradient methods for solving such problems within a high-dimensional framework that allows the data dimension d to grow with (and possibly exceed) the sample size n. This high-dimensional structure precludes the usual global assumptions—namely, strong convexity and smoothness conditions—that underlie classical optimization analysis. We define appropriately restricted versions of these conditions, and show that they are satisfied with high probability for various statistical models. Under these conditions, our theory guarantees that Nesterov's first-order method [12] has a globally geometric rate of convergence up to the statistical precision of the model, meaning the typical Euclidean distance between the true unknown parameter θ^* and the optimal solution $\widehat{\theta}$. This globally linear rate is substantially faster than previous analyses of global convergence for specific methods that yielded only sublinear rates. Our analysis applies to a wide range of M-estimators and statistical models, including sparse linear regression using Lasso (ℓ_1-regularized regression), group Lasso, block sparsity, and low-rank matrix recovery using nuclear norm regularization. Overall, this result reveals an interesting connection between statistical precision 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 consistency and rates for various estimators under high-dimensional scaling, meaning that the data dimension d and other structural parameters are allowed to grow with the sample size n. These results typically involve some assumption regarding the underlying structure of the parameter space, including sparse vectors, low-rank matrices, or structured regression functions, as well as some regularity conditions on 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 programming (Lasso), second-order cone programs for sparse non-parametric regression, and semidefinite programming relaxations for low-rank matrix recovery.

In parallel, a line of recent work (e.g., [3, 7, 6, 5, 12, 18]) focuses on polynomial-time algorithms for solving these types of convex programs. Several authors [2, 6, 1] have used variants of Nesterov's accelerated gradient method [12] to obtain algorithms with a global sublinear rate of convergence. For the special case of compressed sensing (sparse regression with incoherent design), some authors have established fast convergence rates in a local sense–once the iterates are close enough to the optimum [3, 5]. Other authors have studied finite convergence of greedy algorithms (e.g., [18]). If an algorithm identifies the support set of the optimal solution, the problem is then effectively reduced to the lower-dimensional subspace, and thus fast convergence can be guaranteed in a local sense. Also in application to compressed sensing, Garg and Khandekar [4] showed that a thresholded gradient algorithm converges rapidly up to some tolerance; we discuss this result in more detail following our Corollary 2 on this special case of sparse linear models.

Unfortunately, for general convex programs with only Lipschitz conditions, the best convergence rates in a global sense using first-order methods are sub-linear. Much faster global rates—in particular, at a linear or geometric rate—can be achieved if global regularity conditions like strong convexity and smoothness are imposed [11]. However, a challenging aspect of statistical estimation in high dimensions is that the underlying optimization problems can never be globally strongly convex when $d > n$ in typical cases (since the $d \times d$ Hessian matrix is rank-deficient), and global smoothness conditions cannot hold when $d/n \to +\infty$.

In this paper, we analyze a simple variant of the composite gradient method due to Nesterov [12] in application to the optimization problems that underlie regularized M-estimators. Our main contribution is to establish a form of global geometric convergence for this algorithm that holds for a broad class of high-dimensional statistical problems. We do so by leveraging the notion of restricted strong convexity, used in recent work by Negahban et al. [8] to derive various bounds on the statistical error in high-dimensional estimation. Our analysis consists of two parts. We first establish that for optimization problems underlying such M-estimators, appropriately modified notions of restricted strong convexity (RSC) and smoothness (RSM) suffice to establish global linear convergence of a first-order method. Our second contribution is to prove that for the iterates generated by our first-order method, these RSC/RSM assumptions do indeed hold with high probability for a broad class of statistical models, among them sparse linear regression, group-sparse regression, matrix completion, and estimation in generalized linear models. We note in passing that our notion of RSC is related to but slightly different than its previous use for bounding statistical error [8], and hence we cannot use these existing results directly.

An interesting aspect of our results is that we establish global geometric convergence only up to the *statistical precision* of the problem, meaning the typical Euclidean distance $\|\widehat{\theta} - \theta^*\|$ between the true parameter θ^* and the estimate $\widehat{\theta}$ obtained by solving the optimization problem. Note that this is very natural from the statistical perspective, since it is the true parameter θ^* itself (as opposed to the solution $\widehat{\theta}$ of the M-estimator) that is of primary interest, and our analysis allows us to approach it as close as is statistically possible. 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 the paper is organized as follows. In the following section, we give a precise description of the M-estimators considered here, provide definitions of restricted strong convexity and smoothness, and their link to the notion of statistical precision. Section 3 gives a statement of our main result, as well as its corollaries when specialized to various statistical models. Section 4 provides some simulation results that confirm the accuracy of our theoretical predictions. Due to space constraints, we refer the reader to the full-length version of our paper for technical details.

2 Problem formulation and optimization algorithm

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 describe the assumptions that underlie our main result.

A class of regularized M-estimators: 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 collection of n observations drawn i.i.d. from $\mathbb{P}$. 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. In order to do so, we consider the regularized M-estimator

$$\widehat{\theta}_{\lambda_n} \in \arg\min_{\theta \in \Omega} \big\{ \mathcal{L}(\theta; Z_1^n) + \lambda_n \mathcal{R}(\theta) \big\}, \tag{1}$$

where $\mathcal{L} : \Omega \times \mathcal{Z}^n \mapsto \mathbb{R}$ is a loss function, and $\mathcal{R} : \Omega \mapsto \mathbb{R}^+$ is a non-negative regularizer on the parameter space. Throughout this paper, we assume that the loss function $\mathcal{L}$ is convex and differentiable, and that the regularizer $\mathcal{R}$ is a norm. In order to assess the quality of an estimate, we measure the error $\|\widehat{\theta}_{\lambda_n} - \theta^*\|$ in some norm induced by an inner product $\langle \cdot, \cdot \rangle$ on the parameter space. Typical choices are the standard Euclidean inner product and ℓ_2-norm for vectors; the trace inner product and the Frobenius norm for matrices; and the $L^2(\mathbb{P})$ inner product and norm for non-parametric regression. As described in more detail in Section 3.2, a variety of estimators—among them the Lasso, structured non-parametric regression in RKHS, and low-rank matrix recovery—can be cast in this form (1). When the data Z_1^n are clear from the context, we frequently use the shorthand $\mathcal{L}(\cdot)$ for $\mathcal{L}(\cdot; Z_1^n)$.

Composite objective minimization: In general, we expect the loss function $\mathcal{L}$ to be differentiable, while the regularizer $\mathcal{R}$ can be non-differentiable. Nesterov [12] proposed a simple first-order method to exploit this type of structure, and our focus is a slight variant of this procedure. In particular, given some initialization $\theta^0 \in \Omega$, consider the update

$$\theta^{t+1} = \arg\min_{\theta \in \mathbb{B}_{\mathcal{R}}(\rho)} \big\{ \langle \nabla \mathcal{L}(\theta^t), \theta \rangle + \lambda_n \mathcal{R}(\theta) + \frac{\gamma_u}{2} \|\theta - \theta^t\|_2^2 \big\}, \quad \text{for } t = 0, 1, 2, \ldots, \tag{2}$$

where $\gamma_u > 0$ is a parameter related to the smoothness of the loss function, and

$$\mathbb{B}_{\mathcal{R}}(\rho) := \big\{ \theta \in \Omega \mid \mathcal{R}(\theta) \leq \rho \big\} \tag{3}$$

is the ball of radius ρ in the norm defined by the regularizer. The only difference from Nesterov's method is the additional constraint $\theta \in \mathbb{B}_{\mathcal{R}}(\rho)$, which is required for control of early iterates in the high-dimensional setting. Parts of our theory apply to arbitrary choices of the radius ρ; for obtaining results that are statistically order-optimal, a setting $\rho = \Theta(\mathcal{R}(\theta^*))$ with $\theta^* \in \mathbb{B}_{\mathcal{R}}(\rho)$ is sufficient, so that fairly conservative upper bounds on $\mathcal{R}(\theta^*)$ are adequate.

Structural conditions in high dimensions: It is known that under global smoothness and strong convexity assumptions, the procedure (2) enjoys a globally geometric convergence rate, meaning that there is some $\alpha \in (0, 1)$ such that $\|\theta^t - \widehat{\theta}\| = \mathcal{O}(\alpha^t)$ for all iterations $t = 0, 1, 2, \ldots$ (e.g., see Theorem 5 in Nesterov [12]). 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. The resulting $d \times d$ Hessian matrix $\nabla^2 \mathcal{L}(\theta; Z_1^n)$ is often a sum of n rank-1 terms and hence rank-degenerate whenever $n < d$. However, as we show in this paper, in order to obtain fast convergence rates for an optimization method, 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 this intuition. Consider the first-order Taylor series expansion of the loss function around the point θ' in the direction of θ:

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') := \mathcal{L}(\theta) - \mathcal{L}(\theta') - \langle \nabla \mathcal{L}(\theta'), \theta - \theta' \rangle. \tag{4}$$

Definition 1 (Restricted strong convexity (RSC)). We say the loss function $\mathcal{L}$ satisfies the RSC condition with strictly positive parameters $(\gamma_\ell, \kappa_\ell, \delta)$ if

$$\mathcal{T}_{\mathcal{L}}(\theta; \theta') \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|^2 - \kappa_\ell \delta^2 \qquad \text{for all } \theta, \theta' \in \mathbb{B}_{\mathcal{R}}(\rho). \tag{5}$$

In order to gain intuition for this definition, first consider the degenerate setting $\delta = \kappa_\ell = 0$. In this case, imposing the condition (5) for all $\theta \in \Omega$ is equivalent to the usual definition of strong convexity on the optimization set. In contrast, when the pair (δ, κ_ℓ) are strictly positive, the condition (5) only applies to a limited set of vectors. In particular, when θ' is set equal to the optimum $\widehat{\theta}$, and we assume that θ belongs to the set

$$\mathbb{C} := \mathbb{B}_{\mathcal{R}}(\rho) \cap \big\{\theta \in \Omega \;\mid\; \|\theta - \widehat{\theta}\|^2 \geq \frac{4\kappa_\ell}{\gamma_\ell}\delta^2\big\},$$

then condition (5) implies that $\mathcal{T}_{\mathcal{L}}(\theta; \widehat{\theta}) \geq \frac{\gamma_\ell}{4}\|\theta - \widehat{\theta}\|^2$ for all $\theta \in \mathbb{C}$. Thus, for any feasible θ that is not too close to the optimum $\widehat{\theta}$, we are guaranteed strong convexity in the direction $\theta - \widehat{\theta}$.

We now specify an analogous notion of restricted smoothness:

Definition 2 (Restricted smoothness (RSM)). We say the loss function $\mathcal{L}$ satisfies the RSM condition with strictly positive parameters $(\gamma_u, \kappa_u, \delta)$ if

$$\mathcal{T}_{\mathcal{L}}(\theta; \widehat{\theta}) \leq \frac{\gamma_u}{2}\|\theta - \widehat{\theta}\|^2 + \kappa_u \delta^2 \qquad \text{for all } \theta \in \mathbb{B}_{\mathcal{R}}(\rho). \tag{6}$$

Note that the tolerance parameter δ is the same as that in the definition (5). The additional term $\kappa_u \delta^2$ is not present in analogous smoothness conditions in the optimization literature, but it is essential in our set-up.

Loss functions and statistical precision: In order for these definitions to be sensible and of practical interest, it remains to clarify two issues. First, for what types of loss function and regularization pairs can we expect RSC/RSM to hold? Second, what is the smallest tolerance δ with which they can hold? Past work by Negahban et al. [8] has introduced the class of *decomposable regularizers*; it includes various regularizers frequently used in M-estimation, among them ℓ_1-norm regularization, block-sparse regularization, nuclear norm regularization, and various combinations of such norms. Negahban et al. [8] showed that versions of RSC with respect to θ^* hold for suitable loss functions combined with a decomposable regularizer. The definition of RSC given here is related but slightly different: instead of control in a neighborhood of the true parameter θ^*, we need control over the iterates of an algorithm approaching the optimum $\widehat{\theta}$. Nonetheless, it can be also be shown that our form of RSC (and also RSM) holds with high probability for decomposable regularizers, and this fact underlies the corollaries stated in Section 3.2.

With regards to the choice of tolerance parameter δ, as our results will clarify, it makes little sense to be concerned with choices that are substantially smaller than the *statistical precision* of the model. There are various ways in which statistical precision can be defined; one natural one is $\epsilon^2_{\text{stat}} := \mathbb{E}[\|\widehat{\theta}_{\lambda_n} - \theta^*\|^2]$, where the expectation is taken over the randomness in the data-dependent loss function.[1] The statistical precision of various M-estimators under high-dimensional scaling are now relatively well understood, and in the sequel, we will encounter various models for which the RSM/RSC conditions hold with tolerance equal to the statistical precision.

3 Global geometric convergence and its consequences

In this section, we first state the main result of our paper, and discuss some of its consequences. We illustrate its application to several statistical models in Section 3.2.

[1] As written, statistical precision also depends on the choice of λ_n, but our theory will involve specific choices of λ_n that are order-optimal.

3.1 Guarantee of geometric convergence

Recall that $\widehat{\theta}_{\lambda_n}$ denotes any optimal solution to the problem (1). Our main theorem guarantees that if the RSC/RSM conditions hold with tolerance δ, then Algorithm (2) is guaranteed to have a geometric rate of convergence up to this tolerance. The theorem statement involves the objective function $\phi(\theta) = \mathcal{L}(\theta) + \lambda_n \mathcal{R}(\theta)$.

Theorem 1 (Geometric convergence)**.** *Suppose that the loss function satisfies conditions (RSC) and (RSM) with a tolerance δ and parameters $(\gamma_\ell, \gamma_u, \kappa_\ell, \kappa_u)$. Then the sequence $\{\theta^t\}_{t=0}^{\infty}$ generated by the updates* (2) *satisfies*

$$\|\theta^t - \widehat{\theta}\|^2 \le c_0 \left(1 - \frac{\gamma_\ell}{4\gamma_u}\right)^t + c_1 \delta^2 \qquad \text{for all } t = 0, 1, 2, \ldots \tag{7}$$

where $c_0 := \frac{2\,(\phi(0) - \phi(\widehat{\theta}))}{\gamma_\ell}$, and $c_1 := \frac{8\gamma_u}{\gamma_\ell^2}\big(\frac{4\gamma_\ell \kappa_\ell}{\gamma_u} + \kappa_u\big)$.

Remarks: Note that the bound (7) consists of two terms: the first term decays exponentially fast with the contraction coefficient $\alpha := 1 - \frac{\gamma_\ell}{4\gamma_u}$. The second term is an additive offset, which becomes relevant only for t large enough such that $\|\theta^t - \widehat{\theta}\|^2 = \mathcal{O}(\delta^2)$. Thus, the result guarantees a globally geometric rate of convergence up to the tolerance parameter δ. Previous work has focused primarily on the case of sparse linear regression. For this special case, certain methods are known to be globally convergent at sublinear rates (e.g., [2]), meaning of the type $\mathcal{O}(1/t^2)$. The geometric rate of convergence guaranteed by Theorem 1 is exponentially faster. Other work on sparse regression [3, 5] has provided geometric rates of convergence that hold once the iterates are close to the optimum. In contrast, Theorem 1 guarantees geometric convergence if the iterates are not too close to the optimum $\widehat{\theta}$.

In Section 3.2, we describe a number of concrete models for which the (RSC) and (RSM) conditions hold with $\delta \asymp \epsilon_{\text{stat}}$, which leads to the following result.

Corollary 1. *Suppose that the loss function satisfies conditions (RSC) and (RSM) with tolerance $\delta = \mathcal{O}(\epsilon_{stat})$ and parameters $(\gamma_\ell, \gamma_u, \kappa_\ell, \kappa_u)$. Then*

$$T = \mathcal{O}\big(\frac{\log(1/\epsilon_{stat})}{\log(4\gamma_u/(4\gamma_u - \gamma_\ell))}\big) \tag{8}$$

steps of the updates (2) *ensures that $\|\theta^T - \theta^*\|^2 = \mathcal{O}(\epsilon_{stat}^2)$.*

In the setting of statistical recovery, since the true parameter θ^* is of primary interest, there is little point to optimizing to a tolerance beyond the statistical precision. To the best of our knowledge, this result—where fast convergence happens when the optimization error is larger than statistical precision—is the first of its type, and makes for an interesting contrast with other local convergence results.

3.2 Consequences for specific statistical models

We now consider the consequences of Theorem 1 for some specific statistical models. In contrast to the previous deterministic results, these corollaries hold with high probability.

Sparse linear regression: First, we consider the case of *sparse least-squares regression.* Given an unknown regression vector $\theta^* \in \mathbb{R}^d$, suppose that we make n i.i.d. observations of the form $y_i = \langle x_i, \theta^* \rangle + w_i$, where w_i is zero-mean noise. For this model, each observation is of the form $Z_i = (x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}$. 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"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \le R_q\big\}. \tag{9}$$

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 $\mathbb{B}_q(R_q)$ 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 usual Lasso program, with the quadratic loss function $\mathcal{L}(\theta; Z_1^n) := \frac{1}{2n}\sum_{i=1}^n (y_i - \langle x_i,\, \theta\rangle)^2$ and the ℓ_1-norm regularizer $\mathcal{R}(\theta) := \|\theta\|_1$. We consider the Lasso in application to a random design model, in which each predictor vector $x_i \sim N(0, \Sigma)$; we assume that $\max_{j=1,\ldots,d} \Sigma_{jj} \leq 1$ for standardization, and that the condition number $\kappa(\Sigma)$ is finite.

Corollary 2 (Sparse vector recovery)**.** *Suppose that the observation noise w_i is zero-mean and sub-Gaussian with parameter σ, and $\theta^* \in \mathbb{B}_q(R_q)$, and we use the Lasso program with $\lambda_n = 2\sigma\sqrt{\frac{\log d}{n}}$. Then there are universal positive constants $c_i, i = 0, 1, 2, 3$ such that with probability at least $1 - \exp(-c_3 n \lambda_n^2)$, the iterates* (2) *with $\rho^2 = \Theta\big(\sigma^2 R_q (\frac{n}{\log d})^{q/2}\big)$ satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq c_0 \left(1 - \frac{c_2}{\kappa(\Sigma)}\right)^t + c_1 \underbrace{\sigma^2 R_q \left(\frac{\log d}{n}\right)^{1-q/2}}_{\epsilon^2_{stat}} \qquad \text{for all } t = 0, 1, 2, \ldots. \tag{10}$$

It is worth noting that the form of statistical error ϵ_{stat} given in bound (10) is known to be minimax optimal up to constant factors [13]. In related work, Garg and Khandekar [4] showed that for the special case of design matrices that satisfy the restricted isometry property (RIP), a thresholded gradient method has geometric convergence up to the tolerance $\|w\|_2/\sqrt{n} \approx \sigma$. However, this tolerance is independent of sample size, and far larger the statistical error ϵ_{stat} if $n > \log d$; moreover, severe conditions like RIP are not needed to ensure fast convergence. In particular, Corollary 2 guarantees guarantees geometric convergence up to ϵ_{stat} for many random matrices that violate RIP. The proof of Corollary 2 involves exploiting some random matrix theory results [14] in order to verify that the RSC/RSM conditions hold with high probability (see the full-length version for details).

Matrix regression with rank constraints: For a pair of matrices $A, B \in \mathbb{R}^{m \times m}$, we use $\langle\!\langle A,\; B \rangle\!\rangle = \text{trace}(A^T B)$ to denote the trace inner product. Suppose that we are given n i.i.d. observations of the form $y_i = \langle\!\langle X_i,\; \Theta^* \rangle\!\rangle + w_i$, where w_i is zero-mean noise with variance σ^2, and $X_i \in \mathbb{R}^{m \times m}$ is an observation matrix. The parameter space is $\Omega = \mathbb{R}^{m \times m}$ and each observation is of the form $Z_i = (X_i, y_i) \in \mathbb{R}^{m \times m} \times \mathbb{R}$. In many contexts, it is natural to assume that Θ^* is exactly or approximately low rank; applications include collaborative filtering and matrix completion [7, 15], compressed sensing [16], and multitask learning [19, 10, 17]. In order to model such behavior, we let $\sigma(\Theta^*) \in \mathbb{R}^m$ denote the vector of singular values of Θ^* (padded with zeros as necessary), and impose the constraint $\sigma(\Theta^*) \in \mathbb{B}_q(R_q)$. We then consider the M-estimator based on the quadratic loss $\mathcal{L}(\Theta; Z_1^n) = \frac{1}{2n}\sum_{i=1}^n (y_i - \langle\!\langle X_i,\; \Theta \rangle\!\rangle)^2$ combined with the nuclear norm $\mathcal{R}(\Theta) = \|\sigma(\Theta)\|_1$ as the regularizer.

Various problems can be cast within this framework of matrix regression:

- *Matrix completion:* In this case, observation y_i is a noisy version of a randomly selected entry $\Theta^*_{a(i),b(i)}$ of the unknown matrix. It is a special case with $X_i = E_{a(i)b(i)}$, the matrix with one in position $(a(i), b(i))$ and zeros elsewhere.
- *Compressed sensing:* In this case, the observation matrices X_i are dense, drawn from some random ensemble, with the simplest being $X_i \in \mathbb{R}^{m \times m}$ with i.i.d. $N(0, 1)$ entries.
- *Multitask regression:* In this case, the matrix Θ^* is likely to be non-square, with the column size m_2 corresponding to the dimension of the response variable, and m_1 to the number of predictors. Imposing a low-rank constraint on Θ^* is equivalent to requiring that the regression vectors (or columns of the matrix) lie close to a lower-dimensional subspace. See the papers [10, 17] for more details on reformulating this problem as an instance of matrix regression.

For each of these problems, it is possible to show that suitable forms of the RSC/RSM conditions will hold with high probability. For the case of matrix completion, the paper [9] establishes a form of RSC useful for controlling statistical error; this argument can be suitably modified to establish related notions of RSC/RSM required for ensuring fast algorithmic convergence. Similar statements apply to the settings of compressed sensing and multi-task regression. For these matrix regression problems, consider the statistical precision

$$\epsilon^2_{\text{mat}} \asymp \begin{cases} R_q\left(\frac{m\log m}{n}\right)^{1-q/2} & \text{for matrix completion} \\ R_q\left(\frac{m}{n}\right)^{1-q/2} & \text{otherwise,} \end{cases}$$

rates that (up to logarithmic factors) are known to be minimax-optimal [9, 17]. As dictated by this statistical theory, the regularization parameter should be chosen as $\lambda_n = c\sigma\sqrt{\frac{m\log m}{n}}$ for matrix completion, and $\lambda_n = c\sigma\sqrt{\frac{m}{n}}$ otherwise, where $c > 0$ is a universal positive constant. The following result applies to matrix regression problems for which the RSC/RSM conditions hold with tolerance $\delta = \epsilon_{\text{stat}}$.

Corollary 3 (Low-rank matrix recovery). *Suppose that $\sigma(\Theta^*) \in \mathbb{B}_q(R_q)$, and the observation noise is zero-mean σ-sub-Gaussian. Then there are universal positive constants c_1, c_2, c_3 such that with probability at least $1 - \exp(-c_3 n\lambda_n^2)$, the iterates* (2) *with $\rho = \Theta\left(\frac{\epsilon_{mat}}{\lambda_n}\right)$ satisfy*

$$|||\Theta_t - \Theta^*|||_F^2 \leq c_0\nu^t + c_1\epsilon^2_{mat} \qquad \textit{for all } t = 0, 1, 2, \ldots.$$

Here the contraction coefficient $\nu \in (0, 1)$ is a universal constant, independent of (n, m, R_q), depending on the parameters (γ_ℓ, γ_u). We refer the reader to the full-length version for specific form taken for different variants of matrix regression.

4 Simulations

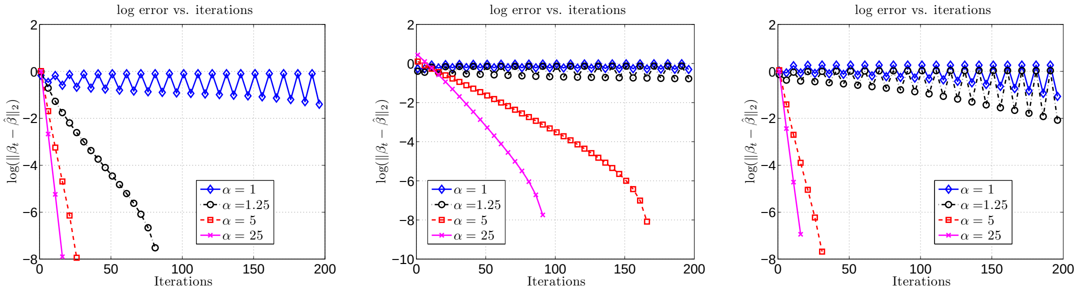
In this section, we provide some experimental results that confirm the accuracy of our theoretical predictions. In particular, these results verify the predicted linear rates of convergence under the conditions of Corollaries 2 and 3.

Sparse regression: We consider a random ensemble of problems, in which each design vector $x_i \in \mathbb{R}^d$ is generated i.i.d. according to the recursion $x(1) = z_1$ and $x(j) = z_j + \upsilon x_i(j-1)$ for $j = 2, \ldots, d$, where the z_j are $\mathcal{N}(0, 1)$, and $\upsilon \in [0, 1)$ is a correlation parameter. The singular values of the resulting covariance matrix Σ satisfy the bounds $\sigma_{\min}(\Sigma) \geq 1/(1+\upsilon)^2$ and $\sigma_{\max}(\Sigma) \leq \frac{2}{(1-\upsilon)^2(1+\upsilon)}$. Note that Σ has a finite condition number for all $\upsilon \in [0, 1)$; for $\upsilon = 0$, it is the identity, but it becomes ill-conditioned as $\upsilon \to 1$. We recall that in this setting $y_i = \langle x_i, \theta^* \rangle + w_i$ where $w_i \sim \mathcal{N}(0, 1)$ and $\theta^* \in \mathbb{B}_q(R_q)$. We study the convergence properties for sample sizes $n = \alpha s \log d$ using different values of α. We note that the per iteration cost of our algorithm is $n \times d$. All our results are averaged over 10 random trials.

Our first experiment is based on taking the correlation parameter $\upsilon = 0$, and the ℓ_q-ball parameter $q = 0$, corresponding to exact sparsity. We then measure convergence rates for $\alpha \in \{1, 1.25, 5, 25\}$ with $d = 40000$ and $s = (\log d)^2$. As shown in Figure 1(a), the procedure fails to converge for $\alpha = 1$: with this setting, the sample size n is too small for conditions (RSC) and (RSM) to hold, so that a constant step size leads to oscillations without these conditions. For α sufficiently large to ensure RSC/RSM, we observe a geometric convergence of the error $\|\theta^t - \widehat{\theta}\|_2$, and the convergence rate is faster for $\alpha = 25$ compared to $\alpha = 5$, since the RSC/RSM constants are better with larger sample size.

On the other hand, we expect the convergence rates to be slower when the condition number of Σ is worse; in addition to address this issue, we ran the same set of experiments with the correlation parameter $\upsilon = 0.5$. As shown in Figure 1(b), in sharp contrast to the case $\upsilon = 0$, we no longer observe geometric convergence for $\alpha = 1.25$, since the conditioning of Σ with $\upsilon = 0.5$ is much poorer than with the identity matrix. Finally, we also expect optimization to be harder as the sparsity parameter $q \in [0, 1]$ is increase away from zero. For larger q, larger sample sizes are required to verify the RSC/RSM conditions. Figure 1(c) shows that even with $\upsilon = 0$, setting $\alpha = 5$ is required for geometric convergence.

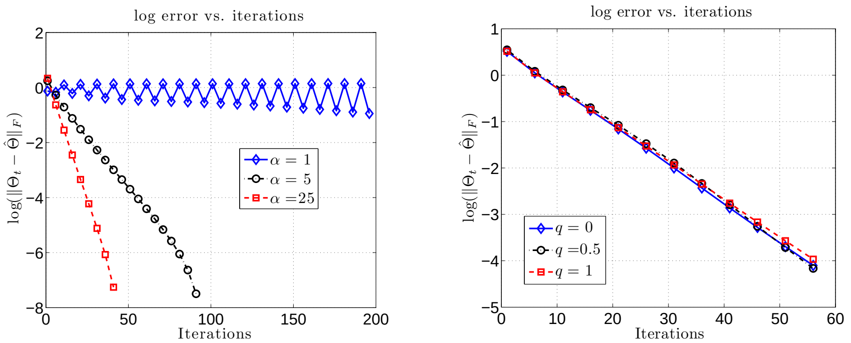
Low-rank matrices: We also performed experiments with two different versions of low-rank matrix regression, each time with $m^2 = 160^2$. The first setting is a version of compressed sensing with matrices $X_i \in \mathbb{R}^{160 \times 160}$ with i.i.d. $N(0, 1)$ entries, and we set $q = 0$,



(a) (b) (c)

Figure 1. Plot of the log of the optimization error $\log(\|\theta^t - \widehat{\theta}\|_2)$ in the sparse linear regression problem. In this problem, $d = 40000$, $s = (\log d)^2$, $n = \alpha s \log d$. Plot (a) shows convergence for the exact sparse case with $q = 0$ and $\Sigma = I$ (i.e. $\upsilon = 0$). In panel (b), we observe how convergence rates change for a non-identity covariance with $\upsilon = 0.5$. Finally plot (c) shows the convergence rates when $\upsilon = 0$, $q = 1$.

and formed a matrix Θ^* with rank $R_0 = \lceil \log m \rceil$. We then performed a series of trials with sample size $n = \alpha R_0\, m$, with the parameter $\alpha \in \{1, 5, 25\}$. The per iteration cost in this case is $n \times m^2$. As seen in Figure 2(a), the general behavior of convergence rates in this problem stays the same as for the sparse linear regression problem: it fails to converge when α is too small, and converges geometrically (with a progressively faster rate) as α increases. Figure 2(b) shows matrix completion also enjoys geometric convergence, for both exactly low-rank ($q = 0$) and approximately low-rank matrices.



(a) (b)

Figure 2. (a) Plot of log Frobenius error $\log(|\!|\!|\Theta^t - \widehat{\Theta}|\!|\!|_F)$ versus number of iterations in matrix compressed sensing for a matrix size $m = 160$ with rank $R_0 = \lceil \log(160) \rceil$, and sample sizes $n = \alpha R_0 m$. For $\alpha = 1$, the algorithm oscillates whereas geometric convergence is obtained for $\alpha \in \{5, 25\}$, consistent with the theoretical prediction. (b) Plot of log Frobenius error $\log(|\!|\!|\Theta^t - \widehat{\Theta}|\!|\!|_F)$ versus number of iterations in matrix completion with approximately low rank matrices ($q \in \{0, 0.5, 1\}$), showing geometric convergence.

5 Discussion

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, and matrix regression with nuclear norm constraints. Overall, our results highlight that the properties of M-estimators favorable for fast rates in a statistical sense can also be used to establish fast rates for optimization algorithms.

Acknowledgements: AA, SN and MJW were partially supported by grants AFOSR-09NL184; SN and MJW acknowledge additional funding from NSF-CDI-0941742.

References

[1] 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.

[2] S. Becker, J. Bobin, and E. J. Candes. Nesta: a fast and accurate first-order method for sparse recovery. Technical report, Stanford University, 2009.

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

[4] R. Garg and R. Khandekar. Gradient descent with sparsification: an iterative algorithm for sparse recovery with restricted isometry property. In *ICML*, New York, NY, USA, 2009. ACM.

[5] 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.

[6] S. Ji and J. Ye. An accelerated gradient method for trace norm minimization. In *International Conference on Machine Learning*, New York, NY, USA, 2009. ACM.

[7] Z. Lin, A. Ganesh, J. Wright, L. Wu, M. Chen, and Y. Ma. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. Technical Report UILU-ENG-09-2214, Univ. Illinois, Urbana-Champaign, July 2009.

[8] 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 Conference*, Vancouver, Canada, December 2009. Full length version arxiv:1010.2731v1.

[9] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. Technical report, UC Berkeley, August 2010. arxiv:1009.2118.

[10] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *Annals of Statistics*, To appear. Originally posted as arxiv:0912.5100.

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

[12] 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.

[13] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. Technical Report arXiv:0910.2042, UC Berkeley, Department of Statistics, 2009.

[14] 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.

[15] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 2010. Posted as arXiv:0910.0651v2.

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

[17] A. Rohde and A. Tsybakov. Estimation of high-dimensional low-rank matrices. Technical Report arXiv:0912.5338v2, Universite de Paris, January 2010.

[18] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, December 2007.

[19] M. Yuan, A. Ekici, Z. Lu, and R. Monteiro. Dimension reduction and coefficient estimation in multivariate linear regression. *Journal Of The Royal Statistical Society Series B*, 69(3):329–346, 2007.