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

Alekh Agarwal[†] Sahand N. Negahban[†] Martin J. Wainwright[⋆,†]
alekh@eecs.berkeley.edu sahand_n@eecs.berkeley.edu wainwrig@stat.berkeley.edu

Department of Statistics[⋆], and
Department of Electrical Engineering and Computer Science[†],
University of California, Berkeley
Berkeley, CA 94720-1776

April 2011

**Abstract**

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 methods for solving such problems, working 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 much of 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 projected gradient descent has a globally geometric rate of convergence up to the *statistical precision* of the model, meaning the typical distance between the true unknown parameter $\theta^*$ and an optimal solution $\widehat{\theta}$. This result is substantially sharper than previous convergence results, which yielded sublinear convergence, or linear convergence only up to the noise level. Our analysis applies to a wide range of $M$-estimators and statistical models, including sparse linear regression using Lasso ($\ell_1$-regularized regression); group Lasso for block sparsity; log-linear models with regularization; low-rank matrix recovery using nuclear norm regularization; and matrix decomposition. Overall, our analysis reveals interesting connections between statistical precision and computational efficiency in high-dimensional estimation.

## 1 Introduction

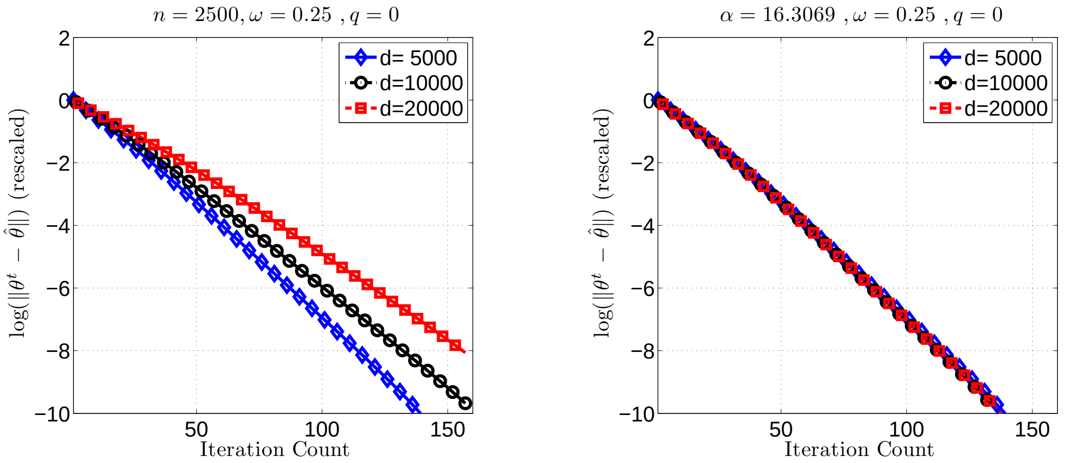
High-dimensional data sets present challenges that are both statistical and computational in nature. On the statistical side, recent years have witnessed a flurry of results on consistency and rates for various estimators under non-asymptotic high-dimensional scaling, meaning that error bounds are provided for general settings of the sample size $n$ and problem dimension $d$, allowing for the possibility that $d \gg n$. These results typically involve some assumption regarding the underlying structure of the parameter space, such as sparse vectors, structured covariance matrices, 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 $\ell_1$-regularized quadratic programs (also known as the Lasso) for sparse linear regression (e.g., [39, 13, 44, 25, 6, 9, 42] and references therein) second-order cone programs (SOCP) for the group Lasso (e.g., [45, 23, 19] and references therein), and semidefinite programming relaxations (SDP) for various problems, including sparse PCA and low-rank matrix estimation (e.g., [11, 35, 38, 2, 37, 29, 36] and references therein).

Many of these programs are instances of convex conic programs, and so can (in principle) be solved to $\epsilon$-accuracy in polynomial time using interior point methods, and other standard methods from convex programming (e.g., see the books [5, 7]). 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., [4, 20, 3]) 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 $\|\cdot\|$. This type of convergence is quite slow, and moreover, it is the best possible with gradient descent-type methods for convex programs with 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 $\kappa^t$, 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 [24] 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 past 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 fast convergence rates in a local sense, meaning guarantees that apply once the iterates are close enough to the optimum [8, 17]. The intuition underlying these results is that once an algorithm identifies the support set of the optimal solution, the problem is then effectively reduced to a lower-dimensional subspace, and thus fast convergence can be guaranteed in a local sense. Also in the setting of compressed sensing, Tropp and Gilbert [40] studied finite convergence of greedy algorithms based on thresholding techniques, and showed linear convergence up to a certain tolerance. For the same class of problems, Garg and Khandekar [16] showed that a thresholded gradient algorithm converges rapidly up to some tolerance. In both of these results, the convergence tolerance is of the order of the noise variance, and hence substantially larger than the true statistical precision of the problem.

The focus of this paper is the convergence rate of the method of projected gradient descent in application to 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. 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. In order to provide intuition for this guarantee, Figure 1 shows the performance of projected gradient descent for a Lasso problem ($\ell_1$-constrained least-squares). In panel (a), we have plotted the logarithm of the optimization error, measured in terms of the Euclidean norm $\|\theta^t - \widehat{\theta}\|$ between the current iterate $\theta^t$ and an optimal solution $\widehat{\theta}$, versus the iteration number $t$. The plot includes three different curves, corresponding to sparse regression problems in dimension $d \in \{5000, 10000, 20000\}$, and a fixed sample size $n = 2500$. Note that all curves are linear (on this logarithmic scale), revealing the geometric convergence predicted by our theory. Moreover, the convergence is geometric even at early iterations, and takes place to a precision far less than the noise level ($\nu^2 = 0.25$ in this example). We also note that the design matrix does not satisfy the restricted isometry property, as assumed in some past work.



(a) (b)

**Figure 1.** Convergence rates of projected gradient descent in application to Lasso programs ($\ell_1$-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.

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—compare the squares for $d = 20000$ to the diamonds for $d = 5000$. This phenomenon reflects the natural intuition that larger problems are, in some sense, "harder" than smaller problems. A notable aspect of our theory is that in addition to guaranteeing geometric convergence, 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 in a certain way according to the dimension $d$ and also other model parameters such as sparsity, then convergence rates should be roughly similar. Panel (b) provides a confirmation of 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 introduced in a recent paper by Negahban et al. [26] for establishing statistical consistency. As we discuss in the sequel, bounding the optimization error introduces new challenges not present when analyzing the statistical error. 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, appropriately modified notions of restricted strong convexity (RSC) and smoothness (RSM) are sufficient to guarantee global linear convergence of projected gradient descent. 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 models, models with group sparsity constraints, and various classes of matrix estimation problems, including matrix completion and matrix decomposition.

An interesting aspect of our results is that the global geometric convergence is not guaranteed to an arbitrary numerical precision, but only to an accuracy related to *statistical precision* of the problem. For a given error norm $\|\cdot\|$, given by the Euclidean or Frobenius norm for most examples in this paper, the statistical precision is given by the mean-squared error $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2]$ between the true parameter $\theta^*$ and the estimate $\widehat{\theta}$ obtained by solving the optimization problem, where the expectation is taken over randomness in the statistical model. Note that this is very natural from the statistical perspective, since it is the true parameter $\theta^*$ 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. Our analysis shows that we can geometrically converge to a parameter $\theta$ 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 a precise formulation of the class of convex programs analyzed in this paper, along with background on the notions of a decomposable regularizer, and properties of the loss function. Section 3 is devoted to the statement of our main convergence result, as well as to the development and discussion of its various corollaries for specific statistical models. In Section 4, we provide a number of empirical results that confirm the sharpness of our theoretical predictions. Finally, Section 5 contains the proofs, with more technical aspects of the arguments deferred to the Appendix.

## 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 projected gradient descent

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. Here the integer $n$ is the *sample size* of the problem. 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 $\Omega$ is some subset of $\mathbb{R}^d$, and the integer $d$ is known as the *ambient dimension* of the problem. In order to measure the "fit" of any given parameter $\theta \in \Omega$ 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 (but not all) 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 under-determined, meaning that the number of observations $n$ is smaller than the ambient dimension $d$. In such settings, without further restrictions on the parameter space $\Omega$, there are various impossibility theorems, asserting that consistent estimates of the unknown parameter $\theta^*$ cannot be obtained. For this reason, it is necessary to assume that the unknown parameter $\theta^*$ either lies within a smaller subset of $\Omega$, or is well-approximated by a member of such subset. In order to incorporate these types of structural constraints, we introduce a *regularizer* $\mathcal{R} : \Omega \to \mathbb{R}_+$ over the parameter space.

With these ingredients, the analysis of this paper applies to optimization problems of the form

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

where $\rho > 0$ is a user-defined radius and

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

where the regularization weight $\lambda_n > 0$ is user-defined. We point out that the radii $\rho$ and $\bar{\rho}$ will 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 $\Omega$, 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. Some remarks on notation: when the radius $\rho$ or the regularization parameter $\lambda_n$ is clear from the context, we will drop the subscript on $\widehat{\theta}$ to ease the notation. Similarly, we frequently adopt the shorthand $\mathcal{L}_n(\theta)$, with the dependence of the loss function on the data being implicitly understood. 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* is applied to the constrained problem (1) while *composite gradient descent* [31] is employed 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 applying the recursive update

$$\theta^{t+1} = \arg\min_{\theta \in \mathbb{B}_{\mathcal{R}}(\rho)} \big\{\mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \theta - \theta^t\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2\big\}, \qquad \text{for } t = 0, 1, 2, \ldots, \tag{3}$$

in the case of projected gradient descent, or the update

$$\theta^{t+1} = \arg\min_{\theta \in \mathbb{B}_{\mathcal{R}}(\bar{\rho})} \big\{ \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t), \, \theta - \theta^t \rangle + \frac{\gamma_u}{2} \|\theta - \theta^t\|^2 + \lambda_n \mathcal{R}(\theta) \big\}, \qquad \text{for } t = 0, 1, 2, \ldots, \tag{4}$$

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 $\theta^{t+1}$ 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 ball $\mathbb{B}_{\mathcal{R}}(\rho) = \{\theta \in \Omega \mid \mathcal{R}(\theta) \leq \rho\}$ of radius $\rho$. In this formulation, we see that the algorithm takes a step in the 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 $\lambda_n$ and $\gamma_u$. 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 $\ell_1$-regularization, it can be obtained by an appropriate form of the soft-thresholding operator.

## 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 [5, 7, 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 $\theta'$ in the direction of $\theta$—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 standard texts for statements and proofs of the result (e.g., Bertsekas [5, Prop. 1.2.3, p. 145], or Nesterov [30, Thm. 2.2.8, p. 88]) for the update (1) and [31] for (2).

[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.

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, 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 $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ 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 $\gamma_\ell$ as the *(lower) curvature parameter*, and to the quantity $\tau_\ell$ as the *tolerance parameter*.

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 $\Omega$. As discussed previously, this type of global strong convexity typically *fails* to hold for high-dimensional inference problems. In contrast, when tolerance parameter $\tau_\ell$ 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, restricted strong convexity imposes a non-trivial constraint only on pairs $\theta \neq \theta'$ for which the inequality (8) holds, and a central part of our analysis will be to prove that, for the sequence of iterates generated by projected gradient descent, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies a constraint of the form (9).

We also observe that the strong convexity of the loss function $\mathcal{L}_n$ also implies the strong convexity of the regularized loss, $\phi_n$ (2) since $\mathcal{R}$ is convex.

We note that this restricted version of strong convexity can be seen as a special instance of the general theory of paraconvexity [32], but we do not know of convergence rates for minimizing general paraconvex functions.

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 parameters $(\gamma_u, \tau_u(\mathcal{L}_n))$ if

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

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 [26]. 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 away 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 $\mathcal{R}$ is $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$-*decomposable* if

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

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

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

and $\Psi(\{0\}) := 0$.

The quantity $\Psi(\overline{\mathcal{M}})$ is a measure of how the two norms differ over the subspace $\overline{\mathcal{M}}$. In particular, it can be seen as the Lipschitz constant of the norm $\mathcal{R}$ with respect to $\|\cdot\|$.

## 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 $\ell_1$-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 $\ell_1$-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 $\mathbb{P}(y_i \mid x_i; \theta^*, \sigma) \propto \exp\left\{\frac{y_i \langle \theta^*, x_i\rangle - \Phi(\langle \theta^*, x_i\rangle)}{c(\sigma)}\right\}$, where $c(\sigma)$ is a known quantity, $\Phi(\cdot)$ is the log-partition function to normalize the density, and $\theta^* \in \mathbb{R}^d$ is an unknown regression vector. In many applications, the regression vector $\theta^*$ is relatively sparse, so that it is natural to impose an $\ell_1$-constraint. Computing the maximum likelihood estimate subject to such a constraint involves solving the convex program[2]

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

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), with the loss function $\mathcal{L}_n(\theta; Z_1^n) = \frac{1}{n} \sum_{i=1}^{n} \big\{ y_i \, \langle \theta, \, x_i \rangle - \Phi(\langle \theta, \, x_i \rangle) \big\}$ and the regularizer $\mathcal{R}(\theta) = \|\theta\|_1 = \sum_{j=1}^{d} |\theta_j|$.

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, 39]. Other forms of log-linear Lasso that are of interest include logistic regression, Poisson regression, and multinomial regression.

**Projected gradient updates:** Computing the gradient of the log-linear loss from equation (13) is straightforward: we have $\nabla \mathcal{L}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} x_i \big\{ y_i - \Phi'(\langle \theta, \, x_i \rangle) \big\}$, 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 $\ell_1$-ball of radius $\rho$. It is well-known that this projection can be characterized in terms of soft-thresholding, and that the projected update (5) can be computed easily. We refer the reader to Duchi et al. [14] for an efficient implementation requiring $\mathcal{O}(d)$ operations.

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

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

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 $\lambda_n$. If the resulting vector has $\ell_1$-norm greater than $\bar{\rho}$, then we project on to the $\ell_1$-ball just like before. Overall, the complexity of the update is still $\mathcal{O}(d)$ as before.

**Decomposability of $\ell_1$-norm:** We now illustrate how the $\ell_1$-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 usual Euclidean inner product) is given by

$$\overline{\mathcal{M}}^{\perp}(S) = \mathcal{M}^{\perp}(S) = \big\{ \beta \in \mathbb{R}^d \;\mid\; \beta_j = 0 \quad \text{for all } j \in S \big\}. \tag{15}$$

[2] The link function $\Phi$ is convex since it is the log-partition function of a canonical exponential family.

To establish the decomposability of the $\ell_1$-norm with respect to the pair $(\mathcal{M}(S), \overline{\mathcal{M}}^{\perp}(S))$, note that any $\alpha \in \mathcal{M}(S)$ can be written in the partitioned form $\alpha = (\alpha_S, 0_{S^c})$, where $\alpha_S \in \mathbb{R}^s$ and $0_{S^c} \in \mathbb{R}^{d-s}$ is a vector of zeros. Similarly, any vector $\beta \in \overline{\mathcal{M}}^{\perp}(S)$ has the partitioned representation $(0_S, \beta_{S^c})$. With these representations, we have the decomposition

$$\|\alpha + \beta\|_1 = \|(\alpha_S, 0) + (0, \beta_{S^c})\|_1 \; = \; \|\alpha\|_1 + \|\beta\|_1.$$

Consequently, for any subset $S$, the $\ell_1$-norm is decomposable with respect to the pairs $(\mathcal{M}(S), \mathcal{M}^{\perp}(S))$.

In analogy to the $\ell_1$-norm, various types of group-sparse norms are also decomposable with respect to non-trivial subspace pairs. We refer the reader to the paper [26] for further discussion and 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, restricted strong convexity for log-linear models requires only that there exist positive numbers $(\gamma_\ell, \tau_\ell(\mathcal{L}_n))$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \Phi''\big(\langle \theta_t, \, x_i \rangle\big) \; \big(\langle x_i, \, \theta - \theta' \rangle\big)^2 \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\theta - \theta') \qquad \text{for all } \theta, \theta' \in \Omega. \tag{16}$$

Restricted smoothness imposes an analogous upper bound on the Taylor error. For a broad class of log-linear models, such bounds hold with with tolerance $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ of the order $\sqrt{\frac{\log d}{n}}$. Further details on such results are provided in the corollaries to follow our main theorem. A detailed discussion of RSC for exponential families in statistical problems is also found in the reference [26].

In the special case of linear regression, we have $\Phi''(t) = 1$ for all $t \in \mathbb{R}$, so that the lower bound (16) 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 $\ell_1$-regularization, the RSC condition is equivalent to the lower bound

$$\frac{\|X(\theta - \theta')\|_2^2}{n} \geq \frac{\gamma_\ell}{2} \|\theta - \theta'\|_2^2 - \tau_\ell(\mathcal{L}_n) \; \|\theta - \theta'\|_1^2 \qquad \text{for all } \theta, \theta' \in \Omega. \tag{17}$$

Such a condition corresponds to a variant of the restricted eigenvalue (RE) conditions that have been studied in the literature [6, 41, 42]. Such RE conditions are significantly milder than the restricted isometry property (RIP); we refer the reader to van de Geer and Buhlmann [42] for an in-depth comparison of different RE conditions. From the results of Raskutti et al. [34], the condition (17) is satisfied with high probability for a broad class of Gaussian random design matrices, 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 $\langle\langle A,\ B\rangle\rangle := \text{trace}(A^T B)$. In order to ease notation, we define $d := \min\{d_1, d_2\}$. 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 a scalar-matrix pair $Z_i = (y_i, X_i) \in \mathbb{R} \times \mathbb{R}^{d_1 \times d_2}$ linked to $\Theta^*$ via the linear model

$$y_i = \langle\langle X_i,\ \Theta^*\rangle\rangle + w_i, \qquad \text{for } i = 1, 2, \ldots, n, \tag{18}$$

where $w_i$ is an additive observation noise. In many contexts, it is natural to assume that $\Theta^*$ 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, 29]) have studied the $M$-estimator

$$\widehat{\Theta} \in \arg \min_{\Theta \in \mathbb{R}^{d_1 \times d_2}} \Big\{ \frac{1}{2n} \sum_{i=1}^{n} \big(y_i - \langle\langle X_i,\ \Theta\rangle\rangle\big)^2 \Big\} \quad \text{such that } |||\Theta|||_1 \leq \rho, \tag{19}$$

or the corresponding regularized version. Here 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. This optimization problem is an instance of a semidefinite program. As discussed in more detail 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 (19), the projected gradient updates take a very simple form—namely

$$\Theta^{t+1} = \Pi\Big(\Theta^t - \frac{1}{\gamma_u} \frac{\sum_{i=1}^n \big(y_i - \langle\langle X_i,\ \Theta^t\rangle\rangle\big) X_i}{n}\Big), \tag{20}$$

where $\Pi$ denotes Euclidean projection onto the nuclear norm ball $\mathbb{B}_1(\rho) = \{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid |||\Theta|||_1 \leq \rho\}$. 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 $\ell_1$-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 2 singular value thresholding operations as was the case for linear regression.

**Decomposability of nuclear norm:** We now define matrix subspaces for which the nuclear norm is decomposable. Given a target matrix $\Theta^*$—that is, a quantity to be estimated—consider its singular value decomposition $\Theta^* = UDV^T$, where the matrix $D \in \mathbb{R}^{d \times d}$ is diagonal, with the ordered singular values of $\Theta^*$ along its diagonal.[3] For an integer $r \in \{1, 2, \ldots, d\}$, let $U^r \in \mathbb{R}^{d \times r}$ denote the matrix formed by the top $r$ left singular vectors of $\Theta^*$ in its columns, and we define the matrix $V^r$ in a similar fashion. Using col to denote the column span of a matrix, we then define the subspaces[4]

$$\mathcal{M}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{col}(\Theta^T) \subseteq \text{col}(V^r),\ \text{col}(\Theta) \subseteq \text{col}(U^r)\big\}, \quad \text{and} \tag{21a}$$

$$\overline{\mathcal{M}}^{\perp}(U^r, V^r) := \big\{\Theta \in \mathbb{R}^{d_1 \times d_2} \mid \text{col}(\Theta^T) \subseteq (\text{col}(V^r))^{\perp},\ \text{col}(\Theta) \subseteq (\text{col}(U^r))^{\perp}\big\}. \tag{21b}$$

[3] Recall our shorthand notation $d = \min\{d_1, d_2\}$.

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

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

# 3 Main result 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).

## 3.1 Geometric convergence

Of primary interest to us in this paper are bounds on the *optimization error* $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, where $\widehat{\theta}$ is any optimal solution to the constrained $M$-estimator (1). For the constraint problem (1) our results apply to a sequence of iterates generated by the projected gradient updates (3), and an associated optimal solution $\widehat{\theta}$ of the program (1) for which the constraint is active, that is $\mathcal{R}(\widehat{\theta}) = \rho$. For the regularized problem (2) our results apply to a sequence of iterates generated by the gradient updates (4) and any optimal solution $\widehat{\theta}$ of the program (2). The statement of our main result also involves the *statistical error* $\Delta^* := \widehat{\theta} - \theta^*$ between the optimum $\widehat{\theta}$ and the nominal parameter $\theta^*$. At a high level, our main result (Theorems 1 and 2) guarantee that under the RSC/RSM conditions, the optimization error $\|\theta^t - \widehat{\theta}\|^2$ 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 of our result is that the contraction occurs only up to a certain tolerance parameter $\epsilon^2$ depending on these same parameters. We first begin with a statement of our claim applied to the iterates (3) followed by a statement applied to equation (4)

We now provide the notation necessary for a precise statement of this claim. Our main result actually involves a family of upper bounds on the optimization error, one for each pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$ of $\mathcal{R}$-decomposable subspaces (see Definition 3). As will be clarified in the sequel, this subspace choice can be optimized for different models so as 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*

$$\kappa(\mathcal{L}_n; \overline{\mathcal{M}}) := \Big\{1 - \frac{\gamma_\ell}{\gamma_u} + \frac{16\Psi^2(\overline{\mathcal{M}})\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\} \Big\{1 - \frac{16\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\gamma_u}\Big\}^{-1}. \tag{22}$$

In addition, we define the *tolerance parameter*

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) := \frac{32\big(\tau_u(\mathcal{L}_n) + \tau_\ell(\mathcal{L}_n)\big)\,\big(2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \Psi(\overline{\mathcal{M}})\|\Delta^*\| + 2\mathcal{R}(\Delta^*)\big)^2}{\gamma_u}, \tag{23}$$

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

In terms of these two ingredients, we now state our 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 $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. Then for any optimum $\widehat{\theta}$ of the problem (1) for which the constraint is active, we have*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})}{1-\kappa} \qquad \textit{for all iterations } t = 0, 1, 2, \ldots. \tag{24}$$

**Remarks:** Theorem 1 actually provides a family of upper bounds, one for each $\mathcal{R}$-decomposable pair $(\mathcal{M}, \overline{\mathcal{M}})$ such that $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. 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 contraction factor $\kappa$ approaches the $1 - \gamma_\ell/\gamma_u$ as the number of samples grows. This is intuitive since the ratio $\gamma_\ell/\gamma_u$ measures the conditioning of the objective—it is essentially a restricted condition number of the Hessian matrix. Hence the result tells us that a well-conditioned problem is easier to minimize than an ill-conditioned one as we would intuitively expect.

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). As shown in its definition (23), this tolerance depends on the choice of decomposable subspaces, the parameters of the RSC/RSM conditions, and the statistical error $\Delta^* = \widehat{\theta} - \theta^*$, corresponding to the difference between the $M$-estimate $\widehat{\theta}$ and the unknown parameter $\theta^*$. In the corollaries of Theorem 1 to follow, we show that it is often the case that the subspaces can 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 tolerance *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.

For future reference, we point out a slight generalization of Theorem 1, used in obtaining some corollaries of our main result. As the proof reveals, it is only necessary to enforce an RSC condition of the form

$$\mathcal{T}_{\mathcal{L}}(\theta^t; \widehat{\theta}) \geq \frac{\gamma_\ell}{2} \, \|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n) \, \mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2, \tag{25}$$

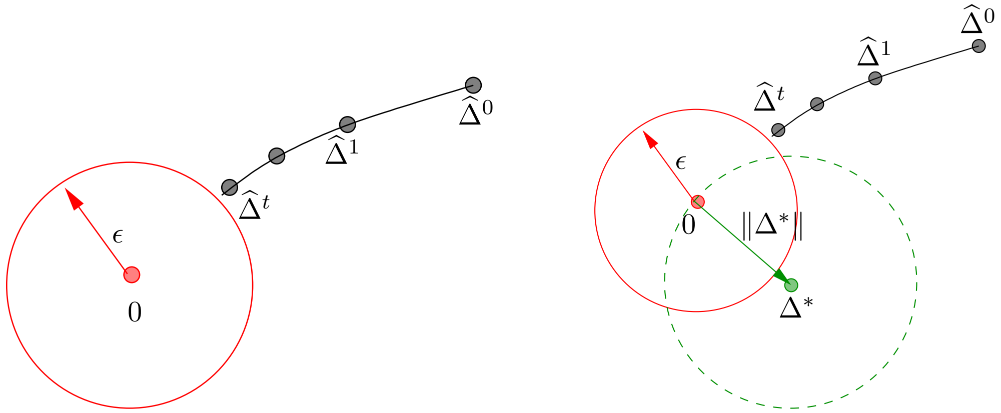
which is milder than the original RSC condition in applying only to differences of the form $\theta^t - \widehat{\theta}$, and allowing for additional slack $\delta$. With this relaxed RSC condition and the same RSM condition as before, our proof shows that

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + \frac{\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) + 2\delta^2/\gamma_u}{1-\kappa} \qquad \text{for all iterations } t = 0, 1, 2, \ldots. \tag{26}$$

We make use of this refinement in the proofs of Corollaries 5 and 6.

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

[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)

**Figure 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 $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \ll \|\Delta^*\|^2$, so that convergence is guaranteed up to a tolerance lower than statistical precision.

term, we can establish that it is of lower order than the statistical precision —namely, the squared difference $\|\widehat{\theta} - \theta^*\|^2$ between an optimal solution $\widehat{\theta}$ to the convex program (1), and the unknown parameter $\theta^*$.

**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*

$$\|\theta^{t+1} - \widehat{\theta}\|^2 \leq \kappa^t \, \|\theta^0 - \widehat{\theta}\|^2 + o\big(\|\widehat{\theta} - \theta^*\|^2\big) \qquad \textit{for all iterations } t = 0, 1, 2, \ldots. \tag{27}$$

Thus, Corollary 1 guarantees that the optimization error decreases geometrically, with contraction factor $\kappa$, 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 will quantify the term $o\big(\|\widehat{\theta} - \theta^*\|^2\big)$ in a more precise manner for different statistical models.

We now present our main result for iterates defined by equation (4). As before, our result will provide a range of bounds indexed by subspace pairs $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ that are $\mathcal{R}$-decomposable. For a fixed pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ such that such that $640\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n) < \gamma_u$, we define the *effective RSC coefficient* as

$$\overline{\gamma_\ell} := \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}}). \tag{28}$$

We also let

$$\xi(\overline{\mathcal{M}}) \;=\; \left(1 - \frac{32\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)^{-1},$$

and

$$\beta(\overline{\mathcal{M}}) \;= 2\left(\frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\tau_\ell(\mathcal{L}_n) + 4\tau_u(\mathcal{L}_n).$$

Finally, define *compound contraction coefficient* as

$$\kappa(\mathcal{L}_n;\overline{\mathcal{M}}) := \left\{1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{128\Psi^2(\overline{\mathcal{M}})\tau_u(\mathcal{L}_n)}{\overline{\gamma_\ell}}\right\}\xi(\overline{\mathcal{M}}) \tag{29}$$

and the *compound tolerance parameter*

$$\epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}}) := 8\,\xi(\overline{\mathcal{M}})\,\beta(\overline{\mathcal{M}})\left(6\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))\right)^2. \tag{30}$$

where in this setting $\Delta^* = \widehat{\theta}_{\lambda_n} - \theta^*$ is the statistical error vector for a specific choice of $\bar{\rho}$ and $\lambda_n$. When the context is clear, we remind the reader that we drop the subscript $\lambda_n$ on the parameter $\widehat{\theta}$. We also require that $\lambda_n$ satisfy two properties, the first stems of statistical requirements [26] and the second to ensure that there is sufficient regularization in order to guarantee RSC and RSM hold:

$$\lambda_n \geq 2\mathcal{R}^*(\nabla\mathcal{L}_n(\theta^*)) \quad \text{and} \tag{31a}$$

$$\lambda_n \geq \frac{16\sqrt{2}}{1-\kappa(\mathcal{L}_n;\overline{\mathcal{M}})}\xi(\overline{\mathcal{M}})\beta(\overline{\mathcal{M}})\bar{\rho} \tag{31b}$$

where we recall that $\bar{\rho}$ is the $\mathcal{R}$ radius of the constraint set in the problem (2). Given these definitions we now state our main theorem on optimization with regularized $M$-estimators.

**Theorem 2.** *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 $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$ be any $\mathcal{R}$-decomposable pair of subspaces such that $\mathcal{M} \subseteq \overline{\mathcal{M}}$ and $0 < \kappa \equiv \kappa(\mathcal{L}_n, \overline{\mathcal{M}}) < 1$. Recall the definition of the composite function $\phi_n(\theta)$ (2) and assume that $\bar{\rho} \geq \mathcal{R}(\theta^*)$ and $\lambda_n$ satisfies conditions* (31). *Then for any $\epsilon \geq \epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})/(1-\kappa)$, we have*

$$\phi_n(\theta^t) - \phi_n(\widehat{\theta}_{\lambda_n}) \leq \epsilon \qquad \textit{for all } t \geq \frac{2\log\frac{\phi_n(\theta^0)-\phi_n(\widehat{\theta}_{\lambda_n})}{\epsilon}}{\log\frac{1}{\kappa}}. \tag{32}$$

This theorem establishes a geometric rate of convergence in terms of the objective values $\phi_n(\theta)$ up to a precision related to the minimax statistical accuracy of the problem. In certain applications, this guarantee in terms of objective values might seem less desirable than a guarantee giving parameter convergence like Theorem 1. However, using the RSC assumption allows us to translate optimization guarantees to guarantees on the parameters as we show next. Indeed, let us assume that under condtions of Theorem 2, we have $\phi_m(\theta) - \phi_n(\widehat{\theta}) \leq \eta$. Then it can be shown that

$$\|\theta - \widehat{\theta}_{\lambda_n}\|^2 \leq \frac{2\eta}{\overline{\gamma_\ell}} + \frac{16\eta^2\tau_\ell(\mathcal{L}_n)}{\overline{\gamma_\ell}\lambda_n^2} + \frac{4\tau_\ell(\mathcal{L}_n)(6\Psi(\overline{\mathcal{M}}) + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)))^2}{\overline{\gamma_\ell}}. \tag{33}$$

Hence, combining Theorem 2 with the above equation, we see that the convergence in parameter values is geometric as well just like previous results, up to an accuracy related to, and smaller than, the minimax statistical precision of the estimation problem. Indeed we can obtain a conclusion similar to 1 for the above theorem as well.

Contrasting the two results, the key upshot of Theorem 2 is in terms of the assumptions needed on $\bar{\rho}$. In order to get a good statistical estimator, the constrained formulation (1) needs a very accurate estimate of $\mathcal{R}(\theta^*)$. On the other hand, the above result only assumes $\bar{\rho} \geq \mathcal{R}(\theta^*)$, which allows much broader settings of $\bar{\rho}$. As for the setting of $\lambda_n$, ways of meeting the condition (31)(a) are well-studied for special cases such as Lasso in the literature [46]. Condition (31)(b) often boils down to a condition on problem parameters such as sample size $n$, $d$, noise variance etc. as we will illustrate in special cases. Hence, despite seemingly having more parameters, the formulation (2) is in some ways computationally more favorable than the constrained version.

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 $\ell_1$-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 $\theta^*$ is sparse. For a parameter $q \in [0,1]$ and radius $R_q > 0$, let us define the $\ell_q$ "ball"

$$\mathbb{B}_q(R_q) := \big\{\theta \in \mathbb{R}^d \mid \sum_{j=1}^{d} |\beta_j|^q \leq R_q\big\}. \tag{34}$$

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 the quadratic loss function $\mathcal{L}(\theta; Z_1^n) := \frac{1}{2n}\|y - X\theta\|_2^2$, 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 $\Sigma$ is a positive definite covariance matrix. We refer to this as the $\Sigma$-*ensemble of random design matrices*, and use $\sigma_{\max}(\Sigma)$ and $\sigma_{\min}(\Sigma)$ to refer the maximum and minimum eigenvalues of $\Sigma$ 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 sub-Gaussian with parameter $\nu^2$. Our convergence rate on the optimization error $\theta^t - \widehat{\theta}$ is stated in terms of the contraction coefficient

$$\kappa := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\} \Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{35}$$

where we have adopted the shorthand

$$\chi_n(\Sigma) := \begin{cases} \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, R_q \left(\frac{\log d}{n}\right)^{1-q/2} & \text{for } q > 0 \\ \frac{c_0\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, s \left(\frac{\log d}{n}\right) & \text{for } q = 0 \end{cases}, \qquad \text{for a numerical constant } c_0, \tag{36}$$

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 [33] from minimax theory on sparse linear regression to be necessary for any method to be statistically consistent over the $\ell_q$-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* $\rho \leq \|\theta^*\|_1$.

*(a)* Exact sparsity: *If* $\theta^*$ *is supported on a subset of cardinality* $s$*, then with probability at least* $1 - \exp(-c_1 \log d)$*, the iterates* (3) *with* $\gamma_u = 2\sigma_{\max}(\Sigma)$ *satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma)\, \|\widehat{\theta} - \theta^*\|_2^2 \qquad \textit{for all } t = 0, 1, 2, \ldots. \tag{37}$$

*(b)* Weak sparsity: *Suppose that* $\theta^* \in \mathbb{B}_q(R_q)$ *for some* $q \in (0,1]$*. Then with probability at least* $1 - \exp(-c_1 \log d)$*, the iterates* (3) *with* $\gamma_u = 2\sigma_{\max}(\Sigma)$ *satisfy*

$$\|\theta^t - \widehat{\theta}\|_2^2 \leq \kappa^t\, \|\theta^0 - \widehat{\theta}\|_2^2 + c_2\, \chi_n(\Sigma) \left\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\widehat{\theta} - \theta^*\|_2^2 \right\}. \tag{38}$$

We provide the proof of Corollary 2 in Section 5.4. Here we 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., [4]), 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 [8, 17], or geometric convergence up to the noise level $\nu^2$ using various methods, including greedy methods [40] 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 $\nu^2$, 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 (Corollary 2(a)) and that of approximate sparsity (Corollary 2(b)). 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 $\ell_q$-ball, and it is no larger than $\|\widehat{\theta} - \theta^*\|_2^2$ with high probability. This assertion follows from known results [33] about minimax rates for linear regression over $\ell_q$-balls, which includes a term of this order.

We can obtain a similar corollary to Theorem 2. We focus only on the exact sparsity for this case, although the result similarly extends to approximate sparsity. We define

$$\overline{\gamma_\ell} = \gamma_\ell - cs\frac{\log d}{n}\zeta(\Sigma).$$
$$\chi_n(\Sigma) = \zeta(\Sigma)s\frac{\log d}{n\overline{\gamma_\ell}}.$$
$$\kappa = \left\{1 - \frac{\sigma_{\min}(\Sigma)}{16\sigma_{\max}(\Sigma)} + c_1\chi_n(\Sigma)\right\}\{1 - c_2\chi_n(\Sigma)\}^{-1},$$

for universal constants $c_1, c_2 > 0$. We also define the critical error

$$\epsilon_0 = \frac{(5 + c_3\chi_n(\Sigma))}{1 - c_2\chi_n(\Sigma)}\zeta(\Sigma)\frac{s\log d}{n}\|\Delta^*\|_2^2, \tag{39}$$

which is $o(\|\Delta^*\|^2)$ if $s\log d = o(n)$. In this case, we have the corollary

**Corollary 3** (Regularized Lasso)**.** *Under conditions of Theorem 2, for all $\epsilon \geq \epsilon_0$, we have*

$$\|\theta^t - \widehat{\theta}_{\lambda_n}\|_2^2 \leq \epsilon \quad \textit{for all } t = \mathcal{O}\left(\frac{\log\frac{\phi_n(\theta^0) - \phi_n(\widehat{\theta}_{\lambda_n})}{\epsilon}}{\log\frac{1}{\kappa}}\right).$$

We also note that the condition (31)(a) can be satisfied by setting $\lambda_n = 2\sqrt{2\nu^2\log d/n}$ w.h.p. Under this setting of $\lambda_n$, condition 31(b) boils down to a condition of the form

$$\nu \geq \frac{c}{1-\kappa}\bar{\rho}\sqrt{\frac{\log d}{n}},$$

which shows an interesting interplay between the noise variance and the constraint radius $\bar{\rho}$. As long $\nu$ is not vanishingly small or $\bar{\rho}$ is not *too large*, this condition will be easily met for $n$ large enough. But if we overestimate $\bar{\rho}$ by too much, geometric convergence might fail.

### 3.3 Matrix regression with rank constraints

We now turn 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 (19). 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 $\sigma(\Theta) = \begin{bmatrix}\sigma_1(\Theta) & \sigma_2(\Theta) & \cdots & \sigma_d(\Theta)\end{bmatrix}$ of singular values. In particular, for a parameter $q \in [0, 1]$, we define the $\ell_q$-"ball" of matrices

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

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, 29]). 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 $\Sigma$, and we define $\zeta_{\text{mat}}(\Sigma) = \sup_{\|u\|_2=1} \sup_{\|v\|_2=1} \operatorname{var}\big(\langle\!\langle X,\, uv^T \rangle\!\rangle\big)$, 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 (20) 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 $\chi_n(\Sigma) := \frac{c_1 \zeta_{\text{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \, R_q \big(\frac{d}{n}\big)^{1-q/2}$ 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 conditions of Theorem 1, consider the semidefinite program* (19) *with* $\rho \leq |\!|\!|\Theta^*|\!|\!|_1$*, and suppose that we apply the projected gradient updates* (20) *with* $\gamma_u = 2\sigma_{\max}(\Sigma)$.

*(a)* Exactly low-rank: *In the case* $q = 0$*, if* $\Theta^*$ *has rank* $r < d$*, then with probability at least* $1 - \exp(-c_0 d)$*, the iterates* (20) *satisfy the bound*

$$|\!|\!|\Theta^t - \widehat{\Theta}|\!|\!|_F^2 \leq \kappa^t |\!|\!|\Theta^0 - \widehat{\Theta}|\!|\!|_F^2 + c_2 \, \chi_n(\Sigma) \, |\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F^2 \qquad \textit{for all } t = 0, 1, 2, \ldots. \tag{42}$$

*(b)* Approximately low-rank: *If* $\Theta^* \in \mathbb{B}_q(R_q)$ *for some* $q \in (0, 1]$*, then with probability at least* $1 - \exp(-c_0 d)$*, the iterates* (20) *satisfy*

$$|\!|\!|\Theta^t - \widehat{\Theta}|\!|\!|_F^2 \leq \kappa^t \, |\!|\!|\Theta^0 - \widehat{\Theta}|\!|\!|_F^2 + c_2 \chi_n(\Sigma) \left\{ R_q \left(\frac{d}{n}\right)^{1-q/2} + |\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F^2 \right\}, \tag{43}$$

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 $|\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F^2$.

A similar Corollary of Theorem 2 can be derived as before and we leave such a developement 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 $\Theta^*$. Applications of this matrix completion problem include collaborative filtering [38], where the rows of the matrix $\Theta^*$ correspond to users, and the columns correspond to items (e.g., movies in the Netflix database), and the entry $\Theta^*_{ab}$ corresponds to user's $a$ rating of item $b$. Given observations of only a subset of the entries of $\Theta^*$, the goal is to fill in, or complete the matrix, thereby making recommendations of movies that a given user has not yet seen.

Matrix completion can be viewed as a particular case of the matrix regression model (18), 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, 38, 28]). Here we consider the $M$-estimator

$$\widehat{\Theta} \in \arg\min_{\Theta\in\Omega} \frac{1}{2n}\sum_{i=1}^{n}\big(y_i - \Theta_{a(i)b(i)}\big)^2 \quad \text{such that } |\!|\!|\Theta|\!|\!|_1 \leq \rho, \tag{44}$$

where $\Omega = \{\Theta \in \mathbb{R}^{d\times d} \mid \|\Theta\|_\infty \leq \frac{\alpha}{d}\}$ is the set of matrices with bounded elementwise $\ell_\infty$ 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 [28], such spikiness control is necessary in order to bound the non-identifiable component of the matrix completion model.

**Corollary 5** (Matrix completion). *Under conditions of Theorem 1, suppose that $\Theta^* \in \mathbb{B}_q(R_q)$, and that we solve the program (44) 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 with probability at least $1 - \exp(-c_1 d\log d)$, there is a contraction coefficient $\kappa_t \in (0,1)$ that decreases with $t$ such that for all iterations $t = 0, 1, 2, \ldots,$*

$$|\!|\!|\Theta^{t+1} - \widehat{\Theta}|\!|\!|_F^2 \leq \kappa_t^t\, |\!|\!|\Theta^0 - \widehat{\Theta}|\!|\!|_F^2 + c_2\Big\{R_q\big(\frac{\alpha^2 d\log d}{n}\big)^{1-q/2} + |\!|\!|\widehat{\Theta} - \Theta^*|\!|\!|_F^2\Big\}. \tag{45}$$

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 31(b) takes the form

$$\lambda_n > \frac{c\alpha\sqrt{d\log d/n}}{1-\kappa},$$

where $\alpha$ is a bound on $\|\theta\|_\infty$. This condition is independent of $\bar{\rho}$, and hence the algorithm always converges geometrically for a large enough setting of $\bar{\rho}$ that satisfies $\bar{\rho} \geq |\!|\!|\Theta^*|\!|\!|_1$, for large enough $n$.

### 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 $\Theta^*$ and $\Gamma^*$, 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 $\Theta^*$ is assumed to be low-rank, and some low-dimensional structural constraint is assumed on the matrix $\Gamma^*$. For example, the papers [12, 10, 18] consider the setting in which $\Gamma^*$ is sparse, while Xu et al. [43] consider a column-sparse model, in which only a few of the columns of $\Gamma^*$ 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 $\ell_1$-norm is decomposable, similar results can easily be derived for the low-rank plus entrywise-sparse setting as well.)

Since $\Theta^*$ 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

$$\|\Gamma\|_{1,2} := \sum_{j=1}^{d_2} \|\Gamma_j\|_2, \tag{46}$$

where $\Gamma_j \in \mathbb{R}^{d_1}$ denotes the $j^{th}$ column of $\Gamma$. 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 $\ell_2$-norm over columns.

In order to estimate the unknown pair $(\Theta^*, \Gamma^*)$, we consider the $M$-estimator

$$(\widehat{\Theta}, \widehat{\Gamma}) := \arg\min_{\Theta,\Gamma} |||Y - \Theta - \Gamma|||_F^2 \quad \text{such that} \quad |||\Theta|||_1 \le \rho_\Theta, \quad \|\Gamma\|_{1,2} \le \rho_\Gamma \text{ and } \|\Theta\|_{\infty,2} \le \frac{\alpha}{\sqrt{d_2}} \tag{47}$$

The first two constraints restrict $\Theta$ and $\Gamma$ to a nuclear norm ball of radius $\rho_\Theta$ and a $(1,2)$-norm ball of radius $\rho_\Gamma$, respectively. The final constraint controls the "spikiness" of the low-rank component $\Theta$, as measured in the $(\infty, 2)$-norm, corresponding to the maximum $\ell_2$-norm over the columns. As with the elementwise $\ell_\infty$-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 $(\Theta^t, \Gamma^t)$ 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 conditions of Theorem 1, suppose that* $\|\Theta^*\|_{\infty,2} \le \frac{\alpha}{\sqrt{d_2}}$ *and* $\Gamma^*$ *has at most* $s$ *non-zero columns. If we solve the convex program* (47) *with* $\rho_\Theta \le |||\Theta^*|||_1$ *and* $\rho_\Gamma \le \|\Gamma^*\|_{1,2}$*, then for all iterations* $t = 0, 1, 2, \ldots$,

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

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). For the matrix decomposition problem, as our analysis shows, the RSC/RSM conditions are guaranteed to hold in a deterministic sense in constrast to the high probability statements in Corollaries 2-5. In this case, 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 the rank of $\Theta^*$ and only depends on the column-sparsity of $\Gamma^*$.

# 4 Simulation results

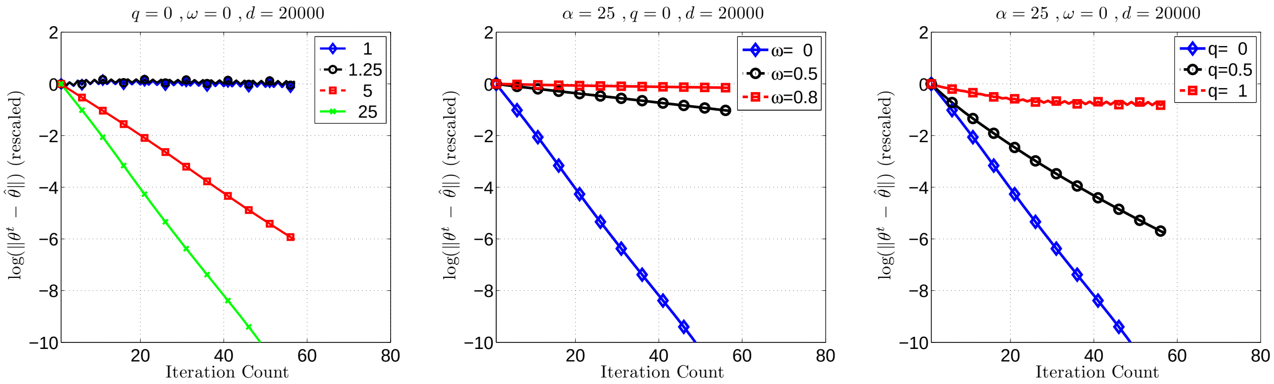
In this section, we provide some experimental results that confirm the accuracy of our theoretical results, in particular showing excellent agreement with the linear rates predicted by our theory. In addition, the rates of convergence slow down for smaller sample sizes, which lead to problems with relatively poor conditioning. In all the simulations reported below, we plot the log error $\|\theta^t - \widehat{\theta}\|$ between the iterate $\theta^t$ 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 begin by considering the linear regression model $y = X\theta^* + w$ where $\theta^*$ 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 $\Sigma = \text{cov}(x_i)$ lie within the interval $[\frac{1}{(1+\omega)^2}, \frac{2}{(1-\omega)^2(1+\omega)}]$, so that $\Sigma$ has a a finite condition number for all $\omega \in [0, 1)$. At one extreme, for $\omega = 0$, the matrix $\Sigma$ is the identity, and so has condition number equal to 1. As $\omega \to 1$, the matrix $\Sigma$ becomes progressively more ill-conditioned, with a condition number that is very large for $\omega$ 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 $\omega$ 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 $\ell_q$-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 $\alpha$ 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.



(a) (b) (c)

**Figure 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 $\omega$ is varied for $q = 0$ and $\alpha = 25$. Plot (c) shows the convergence rates when $\omega = 0$, $\alpha = 25$ and $q$ is varied.

On the other hand, Corollary 2 also predicts that convergence rates should be slower when the condition number of $\Sigma$ 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.

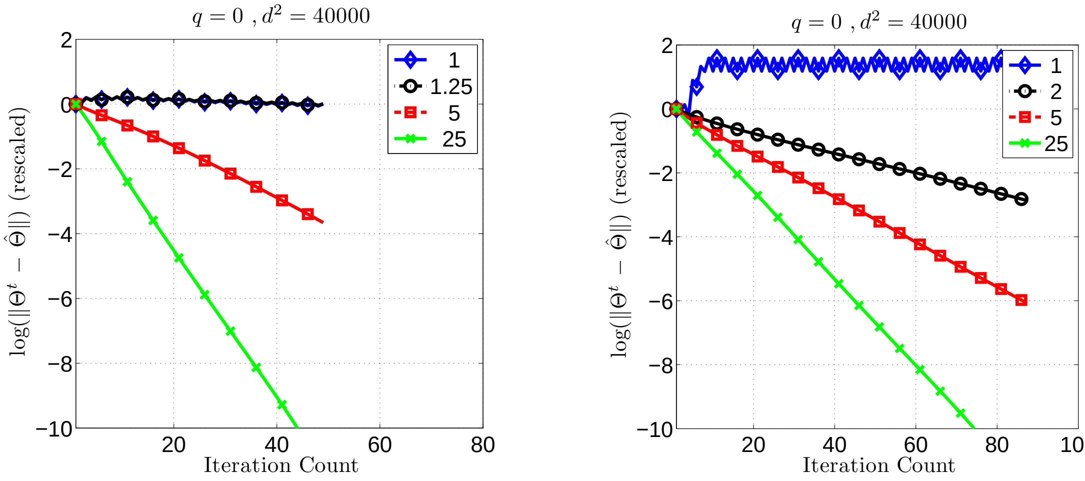
A third prediction of Corollary 2 is that the convergence of projected gradient descent should become slower as the sparsity parameter $q$ is varied between exact sparsity ($q = 0$), and the least sparse case ($q = 1$). (In particular, note for $n > \log d$, the quantity $\chi_n$ from equation (36) is monotonically increasing with $q$.) Panel (c) of Figure 3 shows convergence rates for the fixed sample size $n = 25 s \log d$ and correlation parameter $\omega = 0$, and with the sparsity parameter $q \in \{0, 0.5, 1.0\}$. As expected, the convergence rate convergence slows down as $q$ increases from 0 to 1.

Corollary 2 further predicts 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. We recall that this phenomenon was indeed demonstrated in Figure 1 in the Introduction.

## 4.2 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 $\Theta^*$ 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 $R_1$.



(a)

(b)

**Figure 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 $\alpha \in \{5, 25\}$, 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 $d = 200$, $R_0 = 5$, 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.

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 $\alpha$) is too small, and converges geometrically with a progressively faster rate as $\alpha$ is increased. We have also observed similar types of scaling as the matrix sparsity parameter is increased from $q = 0$ to $q = 1$.

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 $n = \alpha\, R_0\, d \log d$ for $\alpha \in \{1, 2, 5, 25\}$. As shown in panel (b) of Figure 4, projected gradient descent for matrix completion also enjoys geometric convergence for $\alpha$ large enough.

# 5 Proofs

In this section, we provide the proofs of our results. Recall that we use $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ to denote the optimization error, and $\Delta^* = \widehat{\theta} - \theta^*$ to denote the statistical error.

## 5.1 Proof of Theorem 1

Recall that Theorem 1 concerns the constrained problem (1). The proof is based on two technical lemmas. The first lemma guarantees that at each iteration $t = 0, 1, 2, \ldots$, the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$ belongs to an interesting constraint set defined by the regularizer.

**Lemma 1.** *Let $\widehat{\theta}$ be any optimum of the constrained problem* (1) *for which $\mathcal{R}(\widehat{\theta}) = \rho$. Then for any iteration $t = 1, 2, \ldots$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ belongs to the set*

$$\mathbb{S}(\mathcal{M}; \overline{\mathcal{M}}; \theta^*) := \Big\{ \Delta \in \Omega \mid \mathcal{R}(\Delta) \leq 2\,\Psi(\overline{\mathcal{M}})\,\|\Delta\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\mathcal{R}(\Delta^*) + \Psi(\overline{\mathcal{M}})\|\Delta^*\| \Big\}. \tag{48}$$

The proof of this lemma, provided in Appendix A.1, exploits the decomposability of the regularizer in an essential way.

The structure of the set (48) takes a simpler form in the special case when $\mathcal{M}$ is chosen to contain $\theta^*$ and $\overline{\mathcal{M}} = \mathcal{M}$. In this case, we have $\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) = 0$, and hence the optimization error $\widehat{\Delta}^t$ satisfies the inequality

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\mathcal{M})\,\big\{\|\widehat{\Delta}^t\| + \|\Delta^*\|\big\} + 2\mathcal{R}(\Delta^*). \tag{49}$$

An inequality of this type, when combined with the definitions of RSC/RSM, allows us to establish the curvature conditions required to prove globally geometric rates of convergence.

We now state a second lemma under the more general (RSC) condition (25):

**Lemma 2.** *Under conditions* (25) *and (RSM), for all $t = 0, 1, 2, \ldots$, we have*

$$\begin{aligned} &\gamma_u\,\langle \theta^t - \theta^{t+1},\, \theta^t - \widehat{\theta}\rangle \\ &\quad\geq \Big\{\frac{\gamma_u}{2}\|\theta^t - \theta^{t+1}\|^2 - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t)\Big\} + \Big\{\frac{\gamma_\ell}{2}\|\theta^t - \widehat{\theta}\|^2 - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \delta^2\Big\}. \end{aligned} \tag{50}$$

The proof of this lemma, provided in Appendix A.2, follows along the lines of the intermediate result within Theorem 2.2.8 of Nesterov [30], but with some care required to handle the additional terms that arise in our weakened forms of strong convexity and smoothness.

Using these auxiliary results, let us now complete the the proof of Theorem 1. We first note the elementary relation

$$\|\theta^{t+1} - \widehat{\theta}\|^2 = \|\theta^t - \widehat{\theta} - \theta^t + \theta^{t+1}\|^2 = \|\theta^t - \widehat{\theta}\|^2 + \|\theta^t - \theta^{t+1}\|^2 - 2\langle \theta^t - \widehat{\theta},\, \theta^t - \theta^{t+1}\rangle. \tag{51}$$

We now use Lemma 2 and the more general form of RSC (25) to control the cross-term, thereby obtaining the upper bound

$$\begin{aligned}\|\theta^{t+1}-\widehat{\theta}\|^2 &\leq \|\theta^t-\widehat{\theta}\|^2 - \frac{\gamma_\ell}{\gamma_u}\|\theta^t-\widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1}-\theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t-\widehat{\theta}) + \frac{2\delta^2}{\gamma_u} \\ &= \big(1-\frac{\gamma_\ell}{\gamma_u}\big)\|\theta^t-\widehat{\theta}\|^2 + \frac{2\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^{t+1}-\theta^t) + \frac{2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\theta^t-\widehat{\theta}) + \frac{2\delta^2}{\gamma_u}.\end{aligned}$$

We now observe that by triangle inequality and the Cauchy-Schwarz inequality,

$$\mathcal{R}^2(\theta^{t+1}-\theta^t) \leq \big(\mathcal{R}(\theta^{t+1}-\widehat{\theta}) + \mathcal{R}(\widehat{\theta}-\theta^t)\big)^2 \;\leq\; 2\mathcal{R}^2(\theta^{t+1}-\widehat{\theta}) + 2\mathcal{R}^2(\theta^t-\widehat{\theta}).$$

Recall the definition of the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$, we have the upper bound

$$\|\widehat{\Delta}^{t+1}\|^2 \leq \big(1-\frac{\gamma_\ell}{\gamma_u}\big)\|\widehat{\Delta}^t\|^2 + \frac{4\tau_u(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \frac{4\tau_u(\mathcal{L}_n)+2\tau_\ell(\mathcal{L}_n)}{\gamma_u}\mathcal{R}^2(\widehat{\Delta}^t) + \frac{2\delta^2}{\gamma_u}. \tag{52}$$

We now apply Lemma 1 to control the terms involving $\mathcal{R}^2$. In terms of squared quantities, the inequality (48) implies that

$$\mathcal{R}^2(\widehat{\Delta}^t) \leq 4\,\Psi^2(\overline{\mathcal{M}}^\perp)\,\|\widehat{\Delta}^t\|^2 + 2\nu^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}}) \qquad \text{for all } t = 0,1,2,\ldots,$$

where we recall that $\Psi^2(\overline{\mathcal{M}}^\perp)$ is the subspace compatibility (12) and $\nu^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})$ accumulates all the residual terms. Applying this bound twice—once for $t$ and once for $t+1$—and substituting into equation (52) yields that $\big\{1-\frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\tau_u(\mathcal{L}_n)}{\gamma_u}\big\}\|\Delta^{t+1}\|^2$ is upper bounded by

$$\Big\{1-\frac{\gamma_\ell}{\gamma_u}+\frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\big(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n)\big)}{\gamma_u}\Big\}\|\Delta^t\|^2 + \frac{16\big(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n)\big)\nu^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})}{\gamma_u} + \frac{2\delta^2}{\gamma_u}.$$

Under the assumptions of Theorem 1, we are guaranteed that $\frac{16\Psi^2(\overline{\mathcal{M}}^\perp)\tau_u(\mathcal{L}_n)}{\gamma_u} < 1/2$, and so we can re-arrange this inequality into the form

$$\|\Delta^{t+1}\|^2 \leq \kappa\,\|\Delta^t\|^2 + \epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u} \tag{53}$$

where $\kappa$ and $\epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}})$ were previously defined in equations (22) and (23) respectively. Iterating this recursion yields

$$\|\Delta^{t+1}\|^2 \leq \kappa^t\,\|\Delta^0\|^2 + \bigg(\epsilon^2(\Delta^*;\mathcal{M},\overline{\mathcal{M}}) + \frac{2\delta^2}{\gamma_u}\bigg)\,\big(\sum_{j=0}^{t}\kappa^j\big).$$

The assumptions of Theorem 1 guarantee that $\kappa \in (0,1)$, so that summing the geometric series yields the claim (24).

### 5.2 Proof of Theorem 2

The Lagrangian version of the optimization program is based on solving the convex program (2). We recall the definition

$$\phi_n(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$$

and let $\Delta^t_\phi = \phi(\theta^t) - \phi(\widehat{\theta})$ be objective optimization error of the function values. We will drop the subscript on $\phi$ in future to ease notation. The proof of this Theorem will also require two technical lemmas. The first lemma is analogous to Lemma 1 and restricts the optimization error to belong to a constrained set. However, we will apply the lemma only when the optimization error $\Delta^t_\phi$ is sufficiently small for all $t \geq T$.

**Lemma 3** (Iterated Cone Bound (ICB))**.** *Let $\widehat{\theta}$ be any optimum of the regularized $M$-estimator* (2)*. Given some $\eta > 0$, suppose that there exists some integer $T > 0$ such that*

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \eta$$

*for all $t \geq T$. Then for any iteration $t \geq T$ and for any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^\perp)$, the optimization error $\widehat{\Delta}^t := \theta^t - \widehat{\theta}$ satisfies*

$$\mathcal{R}(\widehat{\Delta}^t) \leq 4\Psi(\overline{\mathcal{M}})\|\widehat{\Delta}^t\| + 6\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\min\left(\frac{\eta}{\lambda_n}, \bar{\rho}\right) \tag{54}$$

With this result in hand, we now present our next lemma that guarantees sufficient decrease of the objective optimization error, $\Delta^t_\phi$. We recall that

$$\kappa(\mathcal{L}_n, \overline{\mathcal{M}}) \;= \left(1 - \frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{32\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\left(1 - \frac{32\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)^{-1}.$$

We also define

$$\xi(\overline{\mathcal{M}}) \;= \left(1 - \frac{32\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)^{-1}$$

and

$$\beta(\overline{\mathcal{M}}) \;= 2\left(\frac{\overline{\gamma_\ell}}{4\gamma_u} + \frac{64\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}\right)\tau_\ell(\mathcal{L}_n) + 4\tau_u(\mathcal{L}_n).$$

In the sequel, we drop the arguments of $\kappa, \xi, \beta$ to ease notation. Finally, let $\bar{\epsilon}_{\text{stat}} := 6\Psi(\overline{\mathcal{M}})\|\Delta^*\| + 8\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$ be the statistical error measured with respect to the $\mathcal{R}$ norm. Given these definitions, we now have the following.

**Lemma 4.** *Fix some $\eta > 0$. Suppose that there exists an integer $T > 0$ such that for all $t \geq T$ $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \eta$. Then, under condition* (25) *and (RSM)*

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T}(\phi(\theta^T) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi\beta(\varepsilon^2 + \bar{\epsilon}^2_{stat}),$$

*where $\varepsilon := 2\min(\eta/\lambda_n, \bar{\rho})$.*

We are now in a position to prove prove our main theorem. Our proof will be based on recursively applying the above Lemma at different time epochs. Epoch $k$ will guarantee $\Delta_{\phi}^{t} \leq \eta_k$ for all future epochs, allowing us to apply Lemma 4 with smaller and smaller values of $\epsilon$ until $\epsilon$ reduces to $\bar{\epsilon}_{\text{stat}}$. Note that for all $t \geq T_0 = 0$, we have no bound on $\eta = \phi(\theta^t) - \phi(\widehat{\theta})$. Therefore, we take $\varepsilon_0 = \bar{\rho}$ which is always guaranteed as $\varepsilon = \min(\eta/\lambda_n, \bar{\rho})$. Then Lemma 4 guarantees that for all $t \geq 0$

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^t(\phi(\theta^0) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi\beta(\bar{\rho}^2 + \bar{\epsilon}_{\text{stat}}^2).$$

Since $\kappa < 1$, for all $t \geq \log(1/\eta_1)/\log(1/\kappa)$, we have

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \eta_1 := \frac{4}{1-\kappa}\xi\beta(\bar{\rho}^2 + \bar{\epsilon}_{\text{stat}}^2).$$

Thus, our first epoch is the set of $t$ such that $T_0 \leq t < T_1$ and we observe that $\eta_1 = \phi(\theta^{T_1}) - \phi(\widehat{\theta}) \leq \frac{8\xi\beta}{1-\kappa}\max(\bar{\rho}^2, \bar{\epsilon}_{\text{stat}}^2)$. We now proceed inductively and let $\eta_k = \phi(\theta^{T_k}) - \phi(\widehat{\theta})$ so that for all $t \geq T_k$, $\phi(\theta^t) - \phi(\widehat{\theta}) \leq \eta_k$. Then, by an application of Lemma 4 we have that

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \kappa^{t-T_k}(\phi(\theta^{T_k}) - \phi(\widehat{\theta})) + \frac{2}{1-\kappa}\xi\beta(\varepsilon_k^2 + \bar{\epsilon}_{\text{stat}}^2).$$

Now, since $\kappa < 1$ by assumption, there exists some $T_{k+1}$ such that for all $t \geq T_{k+1}$

$$\phi(\theta^t) - \phi(\widehat{\theta}) \leq \frac{8}{1-\kappa}\xi\beta\max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2). \tag{55}$$

Therefore, we obtain the following recursion:

$$\eta_{k+1} \leq \frac{8}{1-\kappa}\xi\beta\max(\varepsilon_k^2, \bar{\epsilon}_{\text{stat}}^2),$$

where we recall that $\varepsilon_k = 2\,\min(\eta_k/\lambda_n, \bar{\rho})$. By our assumption on $\lambda_n$, we have that $\eta_1/\lambda_n \leq \bar{\rho}/2\sqrt{2}$, so that $\varepsilon_1 \leq \varepsilon_0 = \bar{\rho}$. Thus we apply Equation (55) with $\varepsilon_1 = 2\eta/\lambda_n$ and assuming $\varepsilon_1 \geq \bar{\epsilon}_{\text{stat}}$ we get

$$\begin{aligned}
\eta_2 &\leq \frac{32\xi\beta\eta_1^2}{(1-\kappa)\lambda_n} \\
&\leq \frac{32\xi\beta\bar{\rho}\eta_1}{(1-\kappa)2\sqrt{2}\lambda_n} \quad \text{since } \tfrac{\eta_1}{\lambda_n} \leq \tfrac{\bar{\rho}}{2\sqrt{2}} \\
&\leq \frac{\eta_1}{2},
\end{aligned}$$

from the setting $\lambda_n$. Now it is easy to proceed inductively and verify that

$$\eta_{k+1} \leq \frac{\eta_k}{2^{2^{k-1}}} \quad \text{and} \quad \frac{\eta_{k+1}}{\lambda_n} \leq \frac{\bar{\rho}}{2^{2^k}\sqrt{2}}. \tag{56}$$

Unfolding the recursion, we get that

$$\eta_{k+1} \leq \frac{\eta_1}{2^{2^k-1}}$$

If we are in the first epoch, the claim of the theorem is straightforward from Equation (55). Otherwise, we use the above bound to solve for $t$ in terms of $\epsilon$ as long as $\epsilon \geq 2\bar{\epsilon}_{\text{stat}}$.

## 5.3 Proof of Corollary 1

In order to prove this claim, we must show that $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, as defined in equation (23), is of order lower than $\mathbb{E}[\|\widehat{\theta} - \theta^*\|^2] = \mathbb{E}[\|\Delta^*\|^2]$. We make use of the following lemma, proved in Appendix C:

**Lemma 5.** *If $\rho \leq \mathcal{R}(\theta^*)$, then for any solution $\widehat{\theta}$ of the constrained problem* (1) *and any $\mathcal{R}$-decomposable subspace pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$, the statistical error $\Delta^* = \widehat{\theta} - \theta^*$ satisfies the inequality*

$$\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\| + \mathcal{R}(\Pi_{\mathcal{M}^{\perp}}(\theta^*)). \tag{57}$$

Using this lemma, we can complete the proof of Corollary 1. Recalling the form (23), under the condition $\theta^* \in \mathcal{M}$, we have

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

Using the assumption $\frac{(\tau_u(\mathcal{L}_n)+\tau_\ell(\mathcal{L}_n))\Psi^2(\overline{\mathcal{M}}^{\perp})}{\gamma_u} = o(1)$, it suffices to show that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$. Since Corollary 1 assumes that $\theta^* \in \mathcal{M}$ and hence that $\Pi_{\mathcal{M}^{\perp}}(\theta^*) = 0$, Lemma 5 implies that $\mathcal{R}(\Delta^*) \leq 2\Psi(\overline{\mathcal{M}}^{\perp})\|\Delta^*\|$, as required.

## 5.4 Proofs of Corollaries 2 and 3

The central challenge in proving this result is verifying that suitable forms of the RSC and RSM conditions hold with sufficiently small parameters $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$.

**Lemma 6.** *Define the maximum variance $\zeta(\Sigma) := \max_{j=1,2,\ldots,d} \Sigma_{jj}$. Under the conditions of Corollary 2, there are universal positive constants $(c_0, c_1)$ such that for all $\Delta \in \mathbb{R}^d$, we have*

$$\frac{\|X\Delta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\Delta\|_2^2 - c_1\zeta(\Sigma)\frac{\log d}{n}\ \|\Delta\|_1^2, \qquad \textit{and} \tag{58a}$$

$$\frac{\|X\Delta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\Delta\|_2^2 + c_1\zeta(\Sigma)\frac{\log d}{n}\ \|\Delta\|_1^2, \tag{58b}$$

*with probability at least* $1 - \exp(-c_0\, n)$.

Note that this lemma implies that the RSC and RSM conditions both hold with high probability, in particular with parameters

$$\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma), \text{ and} \qquad \tau_\ell(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n}, \qquad \text{for RSC, and}$$

$$\gamma_u = 2\sigma_{\max}(\Sigma) \text{ and} \qquad \tau_u(\mathcal{L}_n) = c_1\zeta(\Sigma)\frac{\log d}{n} \qquad \text{for RSM.}$$

This lemma has been proved by Raskutti et al. [34] for obtaining minimax rates in sparse linear regression.

Let us first prove Corollary 2 in the special case of hard sparsity ($q = 0$), in which $\theta^*$ is supported on a subset $S$ of cardinality $s$. Let us define the model subspace $\mathcal{M} := \big\{\theta \in \mathbb{R}^d \mid \theta_j = 0 \quad \text{for all } j \notin S\big\}$, so that $\theta^* \in \mathcal{M}$. Recall from Section 2.4.1 that the $\ell_1$-norm is decomposable with respect to $\mathcal{M}$

and $\mathcal{M}^\perp$; as a consequence, we may also set $\overline{\mathcal{M}}^\perp = \mathcal{M}$ in the definitions (22) and (23). By definition (12) of the subspace compatibility between with $\ell_1$-norm as the regularizer, and $\ell_2$-norm as the error norm, we have $\Psi^2(\mathcal{M}) = s$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 6 and substituting into equation (22), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\} \Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{59}$$

where $\chi_n(\Sigma) := \frac{c_2\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, \frac{s \log d}{n}$ for some universal constant $c_2$. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3 \; \phi(\Sigma; s, d, n) \Big\{ \frac{\|\Delta^*\|_1^2}{s} + \|\Delta^*\|_2^2 \Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq \|\theta^*\|_1$, then Lemma 5 (as exploited in the proof of Corollary 1) shows that $\|\Delta^*\|_1^2 \leq 4s\|\Delta^*\|_2^2$, and hence that

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3 \; \chi_n(\Sigma) \; \|\Delta^*\|_2^2.$$

This completes the proof of the claim (37) for $q = 0$.

We now turn to the case $q \in (0, 1]$, for which we bound the term $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. For a truncation level $\mu > 0$ to be chosen, define the set $S_\mu := \big\{j \in \{1, 2, \ldots, d\} \mid |\theta_j^*| > \mu\big\}$, and define the associated subspaces $\mathcal{M} = \mathcal{M}(S_\mu)$ and $\overline{\mathcal{M}}^\perp = \mathcal{M}^\perp(S_\mu)$. By combining Lemma 5 and the definition (23) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \mathcal{M}^\perp(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, \frac{\log d}{n} \big(\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 + \sqrt{|S_\mu|}\, \|\Delta^*\|_2\big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^\perp$ on $S_\mu$. We now choose the threshold $\mu$ optimally, so as to trade-off the term $\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1$, which decreases as $\mu$ increases, with the term $\sqrt{S_\mu}\|\Delta^*\|_2$, which increases as $\mu$ increases.

By definition of $\mathcal{M}^\perp(S_\mu)$, we have

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 = \sum_{j \notin S_\mu} |\theta_j^*| \; = \; \mu \sum_{j \notin S_\mu} \frac{|\theta_j^*|}{\mu} \; \leq \; \mu \sum_{j \notin S_\mu} \Big(\frac{|\theta_j^*|}{\mu}\Big)^q,$$

where the inequality holds since $|\theta_j^*| \leq \mu$ for all $j \notin S_\mu$. Now since $\theta^* \in \mathbb{B}_q(R_q)$, we conclude that

$$\|\Pi_{\mathcal{M}^\perp}(\theta^*)\|_1 \leq \mu^{1-q} \sum_{j \notin S_\mu} |\theta_j^*|^q \; \leq \mu^{1-q} R_q. \tag{60}$$

On the other hand, again using the inclusion $\theta^* \in \mathbb{B}_q(R_q)$, we have $R_q \geq \sum_{j \in S_\mu} |\theta_j^*|^q \; \geq \; |S_\mu|\, \mu^q$ which implies that $|S_\mu| \leq \mu^{-q} R_q$. By combining this bound with inequality (60), we obtain the upper bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, \frac{\log d}{n} \big(\mu^{2-2q} R_q^2 + \mu^{-q} R_q \|\Delta^*\|_2^2\big) \; = \; \frac{c\,\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} \, \frac{\log d}{n} \mu^{-q} R_q \big(\mu^{2-q} R_q + \|\Delta^*\|_2^2\big).$$

Setting $\mu^2 = \frac{\log d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \mathcal{M}^\perp) \leq \varphi_n(\Sigma) \Big\{ R_q \big(\frac{\log d}{n}\big)^{1-q/2} + \|\Delta^*\|_2^2 \Big\}.$$

where $\varphi_n(\Sigma) := \frac{c\zeta(\Sigma)}{\sigma_{\max}(\Sigma)} R_q \big(\frac{\log d}{n}\big)^{1-q/2}$.

Finally, let us verify the stated form of the contraction coefficient. For the given subspace $\overline{\mathcal{M}}^\perp = \mathcal{M}(S_\mu)$ and choice of $\mu$, we have $\Psi^2(\overline{\mathcal{M}}^\perp) = |S_\mu| \leq \mu^{-q} R_q$. From Lemma 6, we have

$$16\Psi^2(\overline{\mathcal{M}}^\perp) \frac{\tau_\ell(\mathcal{L}_n) + \tau_u(\mathcal{L}_n)}{\gamma_u} \leq \varphi_n(\Sigma),$$

and hence, by definition (22) of the contraction coefficient,

$$\kappa \leq \Big\{ 1 - \frac{\gamma_\ell}{2\gamma_u} + \varphi_n(\Sigma) \Big\} \Big\{ 1 - \varphi_n(\Sigma) \Big\}^{-1}.$$

For proving Corollary 3, we observe that the stated settings $\overline{\gamma_\ell}$, $\chi_n(\Sigma)$ and $\kappa$ follow directly from Lemma 6. The bound for condition 2(a) follows from a standard argument about the suprema of $d$ independent Gaussians with variance $\nu$.

### 5.5 Proof of Corollary 4

This proof is analogous to that of Corollary 2, but appropriately adapted to the matrix setting. We first state a lemma that allows us to establish appropriate forms of the RSC/RSM conditions. Recall that we are studying an instance of matrix regression with random design, where the vectorized form $\text{vec}(X)$ of each matrix is drawn from a $N(0, \Sigma)$ distribution, where $\Sigma \in \mathbb{R}^{d^2 \times d^2}$ is some covariance matrix. In order to state this result, let us define the quantity

$$\zeta_{\text{mat}}(\Sigma) := \sup_{\|u\|_2 = 1, \, \|v\|_2 = 1} \text{var}(u^T X v), \quad \text{where } \text{vec}(X) \sim N(0, \Sigma). \tag{61}$$

**Lemma 7.** *Under the conditions of Corollary 4, there are universal positive constants* $(c_0, c_1)$ *such that*

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \geq \frac{1}{2}\, \sigma_{\min}(\Sigma)\, \|\!|\Delta\|\!|_F^2 - c_1 \zeta_{mat}(\Sigma) \frac{d}{n}\, \|\!|\Delta\|\!|_1^2, \qquad \textit{and} \tag{62a}$$

$$\frac{\|\mathfrak{X}_n(\Delta)\|_2^2}{n} \leq 2\, \sigma_{\max}(\Sigma)\, \|\!|\Delta\|\!|_F^2 - c_1\, \zeta_{mat}(\Sigma)\, \frac{d}{n}\, \|\!|\Delta\|\!|_1^2, \qquad \textit{for all } \Delta \in \mathbb{R}^{d \times d}. \tag{62b}$$

*with probability at least* $1 - \exp(-c_0\, n)$.

Given the quadratic nature of the least-squares loss, the bound (62a) implies that the RSC condition holds with $\gamma_\ell = \frac{1}{2}\sigma_{\min}(\Sigma)$ and $\tau_\ell(\mathcal{L}_n) = c_1 \zeta_{\text{mat}}(\Sigma) \frac{d}{n}$, whereas the bound (62b) implies that the RSM condition holds with $\gamma_u = 2\sigma_{\max}(\Sigma)$ and $\tau_u(\mathcal{L}_n) = c_1 \zeta_{\text{mat}}(\Sigma) \frac{d}{n}$.

We now prove Corollary 4 in the special case of exactly low rank matrices ($q = 0$), in which $\Theta^*$ has some rank $r \leq d$. Given the singular value decomposition $\Theta^* = UDV^T$, let $U^r$ and $V^r$ be the $d \times r$ matrices whose columns correspond to the $r$ non-zero (left and right, respectively) singular vectors of $\Theta^*$. As in Section 2.4.2, define the subspace of matrices

$$\mathcal{M}(U^r, V^r) := \big\{ \Theta \in \mathbb{R}^{d \times d} \mid \text{col}(\Theta) \subseteq U^r \text{ and } \text{row}(\Theta) \subseteq V^r \big\}, \tag{63}$$

as well as the associated set $\overline{\mathcal{M}}^{\perp}(U^r, V^r)$. Note that $\Theta^* \in \mathcal{M}$ by construction, and moreover (as discussed in Section 2.4.2, the nuclear norm is decomposable with respect to the pair $(\mathcal{M}, \overline{\mathcal{M}}^{\perp})$.

By definition (12) of the subspace compatibility with nuclear norm as the regularizer and Frobenius norm as the error norm, we have $\Psi^2(\mathcal{M}) = r$. Using the settings of $\tau_\ell(\mathcal{L}_n)$ and $\tau_u(\mathcal{L}_n)$ guaranteed by Lemma 7 and substituting into equation (22), we obtain a contraction coefficient

$$\kappa(\Sigma) := \Big\{1 - \frac{\sigma_{\min}(\Sigma)}{4\sigma_{\max}(\Sigma)} + \chi_n(\Sigma)\Big\} \Big\{1 - \chi_n(\Sigma)\Big\}^{-1}, \tag{64}$$

where $\chi_n(\Sigma) := \frac{c_2 \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)} \frac{rd}{n}$ for some universal constant $c_2$. A similar calculation shows that the tolerance term takes the form

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\ \phi(\Sigma; r, d, n) \Big\{ \frac{|\!|\!|\Delta^*|\!|\!|_1^2}{r} + |\!|\!|\Delta^*|\!|\!|_F^2 \Big\} \qquad \text{for some constant } c_3.$$

Since $\rho \leq |\!|\!|\Theta^*|\!|\!|_1$ by assumption, Lemma 5 (as exploited in the proof of Corollary 1) shows that $|\!|\!|\Delta^*|\!|\!|_1^2 \leq 4r |\!|\!|\Delta^*|\!|\!|_F^2$, and hence that

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}) \leq c_3\ \chi_n(\Sigma)\ |\!|\!|\Delta^*|\!|\!|_F^2,$$

which show the claim (42) for $q = 0$.

We now turn to the case $q \in (0, 1]$; as in the proof of this case for Corollary 2, we bound $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$ using a slightly different choice of the subspace pair. Recall our notation $\sigma_1(\Theta^*) \geq \sigma_2(\Theta^*) \geq \cdots \geq \sigma_d(\Theta^*) \geq 0$ for the ordered singular values of $\Theta^*$. For a threshold $\mu$ to be chosen, define $S_\mu = \big\{j \in \{1, 2, \ldots, d\} \mid \sigma_j(\Theta^*) > \mu\big\}$, and $U(S_\mu) \in \mathbb{R}^{d \times |S_\mu|}$ be the matrix of left singular vectors indexed by $S_\mu$, with the matrix $V(S_\mu)$ defined similarly. We then define the subspace $\mathcal{M}(S_\mu) := \mathcal{M}(U(S_\mu), V(S_\mu))$ in an analogous fashion to equation (63), as well as the subspace $\overline{\mathcal{M}}^{\perp}(S_\mu)$.

Now by a combination of Lemma 5 and the definition (23) of $\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}})$, for any pair $(\mathcal{M}(S_\mu), \overline{\mathcal{M}}^{\perp}(S_\mu))$, we have

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^{\perp}) \leq \frac{c\, \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)}\, \frac{d}{n} \big( \sum_{j \notin S_\mu} \sigma_j(\Theta^*) + \sqrt{|S_\mu|}\ |\!|\!|\Delta^*|\!|\!|_F \big)^2,$$

where to simplify notation, we have omitted the dependence of $\mathcal{M}$ and $\mathcal{M}^{\perp}$ on $S_\mu$. As in the proof of Corollary 2, we now choose the threshold $\mu$ optimally, so as to trade-off the term $\sum_{j \notin S_\mu} \sigma_j(\Theta^*)$ with its competitor $\sqrt{|S_\mu|}\ |\!|\!|\Delta^*|\!|\!|_F$. Exploiting the fact that $\Theta^* \in \mathbb{B}_q(R_q)$ and following the same steps as the proof of Corollary 2 yields the bound

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^{\perp}) \leq \frac{c\, \zeta_{\mathrm{mat}}(\Sigma)}{\sigma_{\max}(\Sigma)}\, \frac{d}{n} \big( \mu^{2-2q} R_q^2 + \mu^{-q} R_q |\!|\!|\Delta^*|\!|\!|_F^2 \big).$$

Setting $\mu^2 = \frac{d}{n}$ then yields

$$\epsilon^2(\Delta^*; \mathcal{M}, \overline{\mathcal{M}}^{\perp}) \leq \varphi_n(\Sigma)\ \Big\{ R_q \big(\frac{d}{n}\big)^{1-q/2} + |\!|\!|\Delta^*|\!|\!|_F^2 \Big\},$$

as claimed. The stated form of the contraction coefficient can be verified by a calculation analogous to the proof of Corollary 2.

### 5.6 Proof of Corollary 5

In this case, we let $\mathfrak{X}_n : \mathbb{R}^{d \times d} \to \mathbb{R}^n$ be the operator defined by the model of random signed matrix sampling [28]. As previously argued, establishing the RSM/RSC property amounts to obtaining a form of uniform control over $\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n}$. More specifically, from the proof of Theorem 1, we see that it suffices to have a form of RSC for the difference $\widehat{\Delta}^t = \Theta^t - \widehat{\Theta}$, and a form of RSM for the difference $\Theta^{t+1} - \Theta^t$. The following two lemmas summarize these claims:

**Lemma 8.** *There is a constant $c$ such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$,*

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2} |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2 - \underbrace{c\alpha \sqrt{\frac{r\, d \log d}{n}} \Big\{ \frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha \sqrt{\frac{r d \log d}{n}} + |\!|\!| \Delta^* |\!|\!|_F \Big\}}_{\delta_\ell(r)}. \tag{65}$$

**Lemma 9.** *There is a constant $c$ such that for all iterations $t = 0, 1, 2, \ldots$ and integers $r = 1, 2, \ldots, d-1$, with probability at least $1 - \exp(-d \log d)$, the difference $\Gamma^t := \Theta^{t+1} - \Theta^t$ satisfies the inequality $\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq 2 |\!|\!| \Gamma^t |\!|\!|_F^2 + \delta_u(r)$, where*

$$\delta_u(r) := c\alpha \sqrt{\frac{r d \log d}{n}} \Big\{ \frac{\sum_{j=r+1}^d \sigma_j(\Theta^*)}{\sqrt{r}} + \alpha \sqrt{\frac{r d \log d}{n}} + |\!|\!| \Delta^* |\!|\!|_F + |\!|\!| \widehat{\Delta}^t |\!|\!|_F + |\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F \Big\}.$$

We now complete the proof of Corollary 5 by a minor modification of the proof of Theorem 1. Recalling the elementary relation (51), we have

$$|\!|\!| \Theta^{t+1} - \widehat{\Theta} |\!|\!|_F^2 = |\!|\!| \Theta^t - \widehat{\Theta} |\!|\!|_F^2 + |\!|\!| \Theta^t - \Theta^{t+1} |\!|\!|_F^2 - 2 \langle\!\langle \Theta^t - \widehat{\Theta},\, \Theta^t - \Theta^{t+1} \rangle\!\rangle.$$

From the proof of Lemma 2, we see that the combination of Lemma 8 and 9 (with $\gamma_\ell = \frac{1}{2}$ and $\gamma_u = 2$) imply that

$$2 \langle\!\langle \Theta^t - \Theta^{t+1},\, \Theta^t - \widehat{\Theta} \rangle\!\rangle \geq |\!|\!| \Theta^t - \Theta^{t+1} |\!|\!|_F^2 + \frac{1}{4} |\!|\!| \Theta^t - \widehat{\Theta} |\!|\!|_F^2 - \delta_u(r) - \delta_\ell(r)$$

and hence that

$$|\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F^2 \leq \frac{3}{4} |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2 + \delta_\ell(r) + \delta_u(r).$$

We substitute the forms of $\delta_\ell(r)$ and $\delta_u(r)$ given in Lemmas 8 and 9 respectively; performing some algebra then yields

$$\Big\{ 1 - \frac{c\,\alpha \sqrt{\frac{rd \log d}{n}}}{|\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F} \Big\} |\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F^2 \leq \Big\{ \frac{3}{4} + \frac{c\alpha \sqrt{\frac{rd \log d}{n}}}{|\!|\!| \widehat{\Delta}^t |\!|\!|_F} \Big\} |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2 + c'\ \delta_\ell(r).$$

Consequently, as long as $\min\{ |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2,\ |\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F^2 \} \geq c_3 \alpha \frac{rd \log d}{n}$ for a sufficiently large constant $c_3$, we are guaranteed the existence of some $\kappa \in (0, 1)$ such that

$$|\!|\!| \widehat{\Delta}^{t+1} |\!|\!|_F^2 \leq \kappa |\!|\!| \widehat{\Delta}^t |\!|\!|_F^2 + c' \delta_\ell(r). \tag{66}$$

Since $\delta_\ell(r) = \Omega(\frac{rd\log d}{n})$, this inequality (66) is valid for all $t = 0, 1, 2, \ldots$ as long as $c'$ is sufficiently large. As in the proof of Theorem 1, iterating the inequality (66) yields

$$|\!|\!|\widehat{\Delta}^{t+1}|\!|\!|_F^2 \leq \kappa^t |\!|\!|\widehat{\Delta}^0|\!|\!|_F^2 + \frac{c'}{1-\kappa}\,\delta_\ell(r). \tag{67}$$

It remains to choose the cut-off $r \in \{1, 2, \ldots, d-1\}$ so as to minimize the term $\delta_\ell(r)$. In particular, when $\Theta^* \in \mathbb{B}_q(R_q)$, then as shown in the paper [29], the optimal choice is $r \asymp \alpha^{-q} R_q \big(\frac{n}{d\log d}\big)^{q/2}$. Substituting into the inequality (67) and performing some algebra yields that there is a universal constant $c_4$ such that the bound

$$|\!|\!|\widehat{\Delta}^{t+1}|\!|\!|_F^2 \leq \kappa^t |\!|\!|\widehat{\Delta}^0|\!|\!|_F^2 + \frac{c_4}{1-\kappa}\Big\{R_q\big(\frac{\alpha d\log d}{n}\big)^{1-q/2} + \sqrt{R_q\big(\frac{\alpha d\log d}{n}\big)^{1-q/2}}\;|\!|\!|\Delta^*|\!|\!|_F\Big\}.$$

holds. Now by the Cauchy-Schwarz inequality we have

$$\sqrt{R_q\big(\frac{\alpha d\log d}{n}\big)^{1-q/2}}\;|\!|\!|\Delta^*|\!|\!|_F \leq \frac{1}{2}R_q\big(\frac{\alpha d\log d}{n}\big)^{1-q/2} + \frac{1}{2}|\!|\!|\Delta^*|\!|\!|_F^2,$$

and the claimed inequality (45) follows.

## 5.7 Proof of Corollary 6

Again the main argument in the proof would be to establish the RSM and RSC properties for the decomposition problem. We define $\widehat{\Delta}_\Theta^t = \Theta^t - \widehat{\Theta}$ and $\widehat{\Delta}_\Gamma^t = \Gamma^t - \widehat{\Gamma}$. We start with giving a lemma that establishes RSC for the differences $(\widehat{\Delta}_\Theta^t, \widehat{\Delta}_\Gamma^t)$. We recall that just like noted in the previous section, it suffices to show RSC only for these differences. Showing RSC/RSM in this example amounts to analyzing $|\!|\!|\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t|\!|\!|_F^2$. We recall that this section assumes that $\Gamma^*$ has only $s$ non-zero columns.

**Lemma 10.** *There is a constant $c$ such that for all iterations $t = 0, 1, 2, \ldots,$*

$$|\!|\!|\widehat{\Delta}_\Theta^t + \widehat{\Delta}_\Gamma^t|\!|\!|_F^2 \geq \frac{1}{2}\big(|\!|\!|\widehat{\Delta}_\Theta^t|\!|\!|_F^2 + |\!|\!|\widehat{\Delta}_\Gamma^t|\!|\!|_F^2\big) - c\alpha\sqrt{\frac{s}{d_2}}\left(|\!|\!|\widehat{\Gamma} - \Gamma^*|\!|\!|_F + \alpha\sqrt{\frac{s}{d_2}}\right) \tag{68}$$

This proof of this lemma follows by a straightforward modification of analogous results in the paper [1].

Matrix decomposition has the interesting property that the RSC condition holds in a deterministic sense (as opposed to with high probability). The same deterministic guarantee holds for the RSM condition; indeed, we have

$$|\!|\!|\widehat{\Delta}_\Delta^t + \widehat{\Delta}_\Gamma^t|\!|\!|_F^2 \leq 2\big(|\!|\!|\widehat{\Delta}_\Theta^t|\!|\!|_F^2 + |\!|\!|\widehat{\Delta}_\Gamma^t|\!|\!|_F^2\big), \tag{69}$$

by Cauchy-Schwartz inequality. Now we appeal to the more general form of Theorem 1 as stated in Equation 26, which gives

$$|\!|\!|\widehat{\Delta}_\Theta^{t+1}|\!|\!|_F^2 + |\!|\!|\widehat{\Delta}_\Gamma^{t+1}|\!|\!|_F^2 \leq \left(\frac{3}{4}\right)^t \big(|\!|\!|\widehat{\Delta}_\Theta^0|\!|\!|_F^2 + |\!|\!|\widehat{\Delta}_\Gamma^0|\!|\!|_F^2\big) + c\sqrt{\frac{\alpha s}{d_2}}\left(|\!|\!|\widehat{\Gamma} - \Gamma^*|\!|\!|_F + \frac{\alpha s}{d_2}\right).$$

The stated form of the corollary follows by an application of Cauchy-Schwarz inequality.

# 6 Discussion

In this paper, we have shown that even though high-dimensional $M$-estimators in statistics are neither strongly convex nor smooth, a simple first-order method 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 $\ell_1$-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. Overall, our results highlight some important connections between computation and statistics: 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:** All three authors were partially supported by grants AFOSR-09NL184; in addition, AA was partially supported by a Microsoft Graduate Fellowship, and SN and MJW acknowledge funding from NSF-CDI-0941742.

# A Auxiliary results for Theorem 1

In this appendix, we provide the proofs of various auxiliary lemmas required in the proof of Theorem 1.

## A.1 Proof of Lemma 1

Since $\theta^t$ and $\widehat{\theta}$ are both feasible and $\widehat{\theta}$ lies on the constraint boundary, we have $\mathcal{R}(\theta^t) \le \mathcal{R}(\widehat{\theta})$. Since $\mathcal{R}(\widehat{\theta}) \le \mathcal{R}(\theta^*) + \mathcal{R}(\widehat{\theta} - \theta^*)$ by triangle inequality, we conclude that

$$\mathcal{R}(\theta^t) \le \mathcal{R}(\theta^*) + \mathcal{R}(\Delta^*).$$

Since $\theta^* = \Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)$, a second application of triangle inequality yields

$$\mathcal{R}(\theta^t) \le \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{70}$$

Now define the difference $\Delta^t := \theta^t - \theta^*$. (Note that this is slightly different from $\widehat{\Delta}^t$, which is measured relative to the optimum $\widehat{\theta}$.) With this notation, we have

$$\begin{aligned}
\mathcal{R}(\theta^t) &= \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\overline{\mathcal{M}}}(\Delta^t) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)\big) \\
&\overset{(i)}{\ge} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}\big(\Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\overline{\mathcal{M}}}(\Delta^t)\big) \\
&\overset{(ii)}{\ge} \mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)\big) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)),
\end{aligned}$$

where steps (i) and (ii) each use the triangle inequality. Now by the decomposability condition, we have $\mathcal{R}\big(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)\big) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t))$, so that we have shown that

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)) \le \mathcal{R}(\theta^t).$$

Combining this inequality with the earlier bound (70) yields

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)) \leq \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*).$$

Re-arranging yields the inequality

$$\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)) \leq \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*). \tag{71}$$

The final step is to translate this inequality into one that applies to the optimization error $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$. Recalling that $\Delta^* = \widehat{\theta} - \theta^*$, we have $\widehat{\Delta}^t = \Delta^t - \Delta^*$, and hence

$$\mathcal{R}(\widehat{\Delta}^t) \leq \mathcal{R}(\Delta^t) + \mathcal{R}(\Delta^*), \qquad \text{by triangle inequality.} \tag{72}$$

In addition, we have

$$\begin{aligned} \mathcal{R}(\Delta^t) \;\leq\; \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^t)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)) &\overset{(i)}{\leq}\; 2\,\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^t)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*) \\ &\overset{(ii)}{\leq} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\overline{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \mathcal{R}(\Delta^*), \end{aligned}$$

where inequality (i) uses the bound (71), and inequality (ii) uses the definition of the subspace compatibility $\Psi$ (12). Combining with the inequality (72) yields

$$\mathcal{R}(\widehat{\Delta}^t) \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\overline{\mathcal{M}}}(\Delta^t)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + 2\mathcal{R}(\Delta^*).$$

Since projection onto a subspace is non-expansive, we have $\|\Pi_{\overline{\mathcal{M}}}(\Delta^t)\| \leq \|\Delta^t\|$, and hence

$$\|\Pi_{\overline{\mathcal{M}}}(\Delta^t)\| \leq \|\widehat{\Delta}^t + \Delta^*\| \;\leq\; \|\widehat{\Delta}^t\| + \|\Delta^*\|.$$

Combining the pieces, we obtain the claim (48).

### A.2 Proof of Lemma 2

We start by applying the RSC assumption to the pair $\widehat{\theta}$ and $\theta^t$, thereby obtaining the lower bound

$$\begin{aligned} \mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^t\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^t\rangle + \langle \nabla\mathcal{L}_n(\theta^t),\, \widehat{\theta} - \theta^{t+1}\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}). \end{aligned} \tag{73}$$

Here the second inequality follows by adding and subtracting terms.

Now for compactness in notation, define $\varphi_t(\theta) := \mathcal{L}_n(\theta^t) + \big\langle \nabla\mathcal{L}_n(\theta^t), \theta - \theta^t\big\rangle + \frac{\gamma_u}{2}\|\theta - \theta^t\|^2$, and note that by definition of the algorithm, the iterate $\theta^{t+1}$ minimizes $\varphi_t(\theta)$ over the ball $\mathbb{B}_{\mathcal{R}}(\rho)$. Moreover, since $\widehat{\theta}$ is feasible, the first-order conditions for optimality imply that $\langle \nabla\varphi_t(\theta^{t+1}),\, \widehat{\theta} - \theta^{t+1}\rangle \geq 0$, or equivalently that $\langle \nabla\mathcal{L}_n(\theta^t) + \gamma_u(\theta^{t+1} - \theta^t),\, \widehat{\theta} - \theta^{t+1}\rangle \;\geq\; 0$. Applying this inequality to the lower bound (73), we find that

$$\begin{aligned} \mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2 &\geq \mathcal{L}_n(\theta^t) + \langle \nabla\mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^t\rangle + \gamma_u\langle \theta^t - \theta^{t+1},\, \widehat{\theta} - \theta^{t+1}\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \varphi_t(\theta^{t+1}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u\langle \theta^t - \theta^{t+1},\, \widehat{\theta} - \theta^{t+1}\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}) \\ &= \varphi_t(\theta^{t+1}) + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u\langle \theta^t - \theta^{t+1},\, \widehat{\theta} - \theta^t\rangle - \tau_\ell(\mathcal{L}_n)\mathcal{R}^2(\theta^t - \widehat{\theta}), \end{aligned} \tag{74}$$

where the last step follows from adding and subtracting $\theta^{t+1}$ in the inner product.

Now by the RSM condition, we have

$$\varphi_t(\theta^{t+1}) \geq \mathcal{L}_n(\theta^{t+1}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) \overset{(a)}{\geq} \mathcal{L}_n(\widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t), \tag{75}$$

where inequality (a) follows by the optimality of $\widehat{\theta}$, and feasibility of $\theta^{t+1}$. Combining this inequality with the previous bound (74) yields that $\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_\ell}{2}\|\widehat{\theta} - \theta^t\|^2$ is lower bounded by

$$\mathcal{L}_n(\widehat{\theta}) - \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \gamma_u\langle\theta^t - \theta^{t+1},\, \widehat{\theta} - \theta^t\rangle - \tau_\ell(\mathcal{L}_n)\,\mathcal{R}^2(\theta^t - \widehat{\theta}) - \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t),$$

and the claim (50) follows by some algebraic re-arranging.

# B Auxiliary results for Theorem 2

We now prove the auxiliary lemmas required in the proof of Theorem 2.

## B.1 Proof of Lemma 3

The proof of the Iterated Cone Bound follows the analogous proof in [26], with some changes to adapt the statement to the optimization setting. In that paper, the cone condition was only established between the estimate $\widehat{\theta}$ and the true set of parameters $\theta^*$, so that by definition $\phi(\widehat{\theta}) \leq \phi(\theta^*)$. In our setting, we wish to establish a similar cone bound between the iterates and $\widehat{\theta}$. However, by our assumption, we do not have an exact inequality; instead, we have that $\phi(\theta^t) \leq \phi(\widehat{\theta}) + \eta$ for all $t \geq T$, which implies that

$$\phi(\theta^t) \;\leq\; \phi(\theta^*) + \eta.$$

Thus, we may fix some $\theta$ that satisfies the above inequality. With that, we can let $\Delta = \theta - \theta^*$, $\widehat{\Delta} = \theta - \widehat{\theta}$, and $\Delta^* = \widehat{\theta} - \theta^*$. We may follow the exact same steps outlined in the paper [26] to obtain

$$\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) \;\leq\; 3\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) + 4\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) + \frac{2\eta}{\lambda_n}. \tag{76}$$

With the above result in hand, we may substitute $\Delta = \widehat{\Delta} - \Delta^*$ in order to rewrite the above equation in terms of $\widehat{\Delta}$. Thus, by simple algebra, we have our desired result. For completeness, we will now establish that equation (76) holds. By assumption, we have that

$$\mathcal{L}_n(\theta^* + \Delta) + \lambda_n\mathcal{R}(\theta^* + \Delta) \;\leq\; \mathcal{L}_n(\theta^*) + \lambda_n\mathcal{R}(\theta^*) + \eta.$$

Subtracting $\big\langle\nabla\mathcal{L}_n\theta^*,\, \Delta\big\rangle$ from each side, we get

$$\mathcal{L}_n(\theta^* + \Delta) - \mathcal{L}_n(\theta^*) - \big\langle\nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle + \lambda_n\mathcal{R}(\theta^* + \Delta) - \lambda_n\mathcal{R}(\theta^*) \;\leq\; -\big\langle\nabla\mathcal{L}_n\theta^*,\, \Delta\big\rangle + \eta.$$

Using convexity of $\mathcal{L}_n$, the above expression simplifies to

$$\lambda_n\mathcal{R}(\theta^* + \Delta) - \lambda_n\mathcal{R}(\theta^*) \;\leq\; -\big\langle\nabla\mathcal{L}_n(\theta^*),\, \Delta\big\rangle + \eta.$$

Applying Hölder's inequality to $\langle \nabla \mathcal{L}_n(\theta^*),\, \Delta \rangle$ in the pair of norms $\mathcal{R}, \mathcal{R}^*$ along with triangle inequality, and using the fact that $\lambda_n \geq 2\mathcal{R}^*(\nabla \mathcal{L}_n \theta^*)$ yields

$$\lambda_n \mathcal{R}(\theta^* + \Delta) - \lambda_n \mathcal{R}(\theta^*) \;\leq\; \frac{\lambda_n}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) + \frac{\lambda_n}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) + \eta,$$

since $\Delta = \Pi_{\overline{\mathcal{M}}}(\Delta) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta)$. Applying the triangle inequality with $\Delta$ and $\theta^* = \Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)$; and dividing by $\lambda_n$ we have

$$\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) - \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) - 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \;\leq\; \frac{1}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) + \frac{1}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) + \frac{\eta}{\lambda_n}.$$

Now, since $\Pi_{\mathcal{M}}(\theta^*) \in \mathcal{M}$ and $\Pi_{\overline{\mathcal{M}}^\perp}(\Delta) \in \overline{\mathcal{M}}^\perp$, by decomposability $\mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta))$ so that

$$\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) - 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \;\leq\; \frac{1}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta)) + \frac{1}{2}\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta)) + \frac{\eta}{\lambda_n}.$$

Rearranging terms yields the desired result.

## B.2 Proof of Lemma 4

The proof of this result follows similar lines as the proof of convergence by Nesterov [31]. We recall that $\phi(\theta) = \mathcal{L}_n(\theta) + \lambda_n \mathcal{R}(\theta)$, $\widehat{\Delta}^t = \theta^t - \widehat{\theta}$, and that $\Delta^t_\phi = \phi(\theta^t) - \phi(\widehat{\theta})$. We note that by the definition of $\theta^{t+1}$ and by RSM, we have for all $\alpha \in (0,1)$

$$\begin{aligned}
\phi(\theta^{t+1}) &\leq \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t),\, \theta^{t+1} - \theta^t \rangle + \frac{\gamma_u}{2}\|\theta^{t+1} - \theta^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + \lambda_n \mathcal{R}(\theta^{t+1}) \\
&\leq \mathcal{L}_n(\theta^t) + \langle \nabla \mathcal{L}_n(\theta^t),\, \alpha\widehat{\theta} + (1-\alpha)\theta^t - \theta^t \rangle + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\theta^{t+1} - \theta^t) + \lambda_n \mathcal{R}(\alpha\widehat{\theta} + (1-\alpha)\theta^t) \\
&\leq \phi(\alpha\widehat{\theta} + (1-\alpha)\theta^t) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\widehat{\Delta}^{t+1} - \widehat{\Delta}^t) \qquad \text{(using convexity of } \phi(\theta)) \\
&\leq \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\widehat{\Delta}^{t+1}) + \tau_u(\mathcal{L}_n)\mathcal{R}^2(\widehat{\Delta}^t).
\end{aligned}$$

Now we note that by Lemma 3

$$\mathcal{R}^2(\widehat{\Delta}^t) \;\leq 16\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 2v^2$$

where $v = \bar{\epsilon}_{\text{stat}}(\mathcal{M}, \overline{\mathcal{M}}) + 2\min(\frac{\eta}{\lambda_n}, \bar{\rho})$, is a constant independent of $\theta^t$. Therefore, we then have

$$\begin{aligned}
\phi(\theta^{t+1}) \leq \phi(\theta^t) - \alpha(\phi(\theta^t) - \phi(\widehat{\theta})) + \frac{\gamma_u \alpha^2}{2}\|\widehat{\Delta}^t\|^2 \qquad &+ \tau_u(\mathcal{L}_n)16\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^{t+1}\|^2 \\
+ \tau_u(\mathcal{L}_n)16\Psi^2(\overline{\mathcal{M}})\|\widehat{\Delta}^t\|^2 + 4\tau_u(\mathcal{L}_n)v^2
\end{aligned}$$

Now, we must translate quantities related to $\widehat{\Delta}^t$ to functional values. In order to do so, we apply RSC at $\theta^t$ and $\theta^{t+1}$ around $\widehat{\theta}$. Using the optimality of $\widehat{\theta}$ and Lemma 3, we get the following simplified form of RSC

$$\|\widehat{\Delta}^t\|^2 \left(\frac{\gamma_\ell}{2} - 32\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\right) \;\le\; \Delta^t_\phi + 2\tau_\ell(\mathcal{L}_n)v^2.$$

We now let $\overline{\gamma_\ell} = \gamma_\ell - 64\tau_\ell(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})$. Therefore, an application of RSC and some algebra yields that

$$\begin{aligned}\phi(\theta^{t+1}) \;&\le \phi\theta^t - \alpha\Delta^t_\phi + \left(\frac{\gamma_u\alpha^2}{\overline{\gamma_\ell}} + 16\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})\right)(\Delta^t_\phi + 2\tau_\ell(\mathcal{L}_n)v^2) + \\ &\qquad \frac{32\tau_u(\mathcal{L}_n)\Psi^2(\overline{\mathcal{M}})}{\overline{\gamma_\ell}}(\Delta^{t+1}_\phi + 2\tau_\ell v^2) + 4\tau_u(\mathcal{L}_n)v^2.\end{aligned}$$

We may let all of the constant terms that are independent of $\theta^t$ to be $g$. With that, by adding and subtracting $\phi(\widehat{\theta})$ from both sides of the above equation and letting $\alpha = \frac{\overline{\gamma_\ell}}{2\gamma_u}$, we have

$$\Delta^{t+1}_\phi \le \kappa\Delta^t_\phi + g.$$

Thus, iterating the above expression yields that

$$\Delta^t_\phi \le \kappa^t\Delta^0_\phi + \frac{g}{1-\kappa},$$

since by assumption $\kappa < 1$, thus establishing our claim after expanding $g$.

# C Proof of Lemma 5

Given the condition $\mathcal{R}(\widehat{\theta}) \le \rho \le \mathcal{R}(\theta^*)$, we have $\mathcal{R}(\widehat{\theta}) = \mathcal{R}(\theta^* + \Delta^*) \le \mathcal{R}(\theta^*)$. By triangle inequality, we have

$$\mathcal{R}(\theta^*) = \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*)) \;\le\; \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)).$$

We then write

$$\begin{aligned}\mathcal{R}(\theta^* + \Delta^*) &= \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\mathcal{M}^\perp}(\theta^*) + \Pi_{\overline{\mathcal{M}}}(\Delta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^*)) \\ &\overset{(i)}{\ge} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*) + \Pi_{\overline{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{=} \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^*)) - \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^*)) - \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where the bound (i) follows by triangle inequality, and step (ii) uses the decomposability of $\mathcal{R}$ over the pair $\mathcal{M}$ and $\overline{\mathcal{M}}^\perp$. By combining this lower bound with the previously established upper bound

$$\mathcal{R}(\theta^* + \Delta^*) \le \mathcal{R}(\Pi_{\mathcal{M}}(\theta^*)) + \mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),$$

we conclude that $\mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^*)) \le \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*))$. Finally, by triangle inequality, we have $\mathcal{R}(\Delta^*) \le \mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^*)) + \mathcal{R}(\Pi_{\overline{\mathcal{M}}^\perp}(\Delta^*))$, and hence

$$\begin{aligned}\mathcal{R}(\Delta^*) &\le 2\mathcal{R}(\Pi_{\overline{\mathcal{M}}}(\Delta^*)) + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(i)}{\le} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Pi_{\overline{\mathcal{M}}}(\Delta^*)\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)) \\ &\overset{(ii)}{\le} 2\,\Psi(\overline{\mathcal{M}}^\perp)\|\Delta^*\| + 2\mathcal{R}(\Pi_{\mathcal{M}^\perp}(\theta^*)),\end{aligned}$$

where inequality (i) follows from Definition (4) of the subspace compatibility $\Psi$, and the bound (ii) follows from non-expansivity of projection onto a subspace.

## D A general result on Gaussian observation operators

In this appendix, we state a general result about a Gaussian random matrices, and show how it can be adapted to prove Lemmas 6 and 7. Let $X \in \mathbb{R}^{n \times d}$ be a Gaussian random matrix with i.i.d. rows $x_i \sim N(0, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is a covariance matrix. We refer to $X$ as a sample from the $\Sigma$-Gaussian ensemble. In order to state the result, we use $\Sigma^{1/2}$ to denote the symmetric matrix square root.

**Proposition 1.** *Given a random matrix $X$ drawn from the $\Sigma$-Gaussian ensemble, there are universal constants $c_i$, $i = 0, 1$ such that*

$$\frac{\|X\theta\|_2^2}{n} \geq \frac{1}{2}\|\Sigma^{1/2}\theta\|_2^2 - c_1 \frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n} \mathcal{R}^2(\theta) \qquad \textit{and} \tag{77a}$$

$$\frac{\|X\theta\|_2^2}{n} \leq 2\|\Sigma^{1/2}\theta\|_2^2 + c_1 \frac{(\mathbb{E}[\mathcal{R}^*(x_i)])^2}{n} \mathcal{R}^2(\theta) \qquad \textit{for all } \theta \in \mathbb{R}^d \tag{77b}$$

*with probability greater than $1 - \exp(-c_0\, n)$.*

We omit the proof of this result. The two special instances proved in Lemma 6 and 7 have been proved in the papers [33] and [27] resp. We now show how Proposition 1 can be used to recover various lemmas required in our proofs.

**Proof of Lemma 6:** We begin by establishing this auxiliary result required in the proof of Corollary 2. When $\mathcal{R}(\cdot) = \|\cdot\|_1$, we have $\mathcal{R}^*(\cdot) = \|\cdot\|_\infty$. Moreover, the random vector $x_i \sim N(0, \Sigma)$ can be written as $x_i = \Sigma^{1/2} w$, where $w \sim N(0, I_{d \times d})$ is standard normal. Consequently, using properties of Gaussian maxima [21] and defining $\zeta(\Sigma) = \max_{j=1,2,\ldots,d} \Sigma_{jj}$, we have the bound

$$(\mathbb{E}[\|x_i\|_\infty])^2 \;\leq\; \zeta(\Sigma)\; (\mathbb{E}[\|w\|_\infty])^2 \;\leq\; 3\zeta(\Sigma)\; \sqrt{\log d}.$$

Substituting into Proposition 1 yields the claims (58a) and (58b).

**Proof of Lemma 7:** In order to prove this claim, we view each random observation matrix $X_i \in \mathbb{R}^{d \times d}$ as a $d = d^2$ vector (namely the quantity $\operatorname{vec}(X_i)$), and apply Proposition 1 in this vectorized setting. Given the standard Gaussian vector $w \in \mathbb{R}^{d^2}$, we let $W \in \mathbb{R}^{d \times d}$ be the random matrix such that $\operatorname{vec}(W) = w$. With this notation, the term $\mathcal{R}^*(\operatorname{vec}(X_i))$ is equivalent to the operator norm $|\!|\!|X_i|\!|\!|_{\text{op}}$. As shown in Negahban and Wainwright [29], $\mathbb{E}[|\!|\!|X_i|\!|\!|_{\text{op}}] \leq 24\zeta_{\text{mat}}(\Sigma)\, \sqrt{d}$, where $\zeta_{\text{mat}}$ was previously defined (61).

## E Auxiliary results for Corollary 5

In this section, we provide the proofs of Lemmas 8 and 9 that play a central role in the proof of Corollary 5. In order to do so, we require the following result, which is a re-statement of a theorem due to Negahban and Wainwright [28]:

**Proposition 2.** *For the matrix completion operator* $\mathfrak{X}_n$*, there are universal positive constants* $(c_1, c_2)$ *such that*

$$\left|\frac{\|\mathfrak{X}_n(\Theta)\|_2^2}{n} - |\!|\!|\Theta|\!|\!|_F^2\right| \leq c_1\, d\|\Theta\|_\infty\; |\!|\!|\Theta|\!|\!|_1 \sqrt{\frac{d\log d}{n}} \;+\; c_2\left(d\|\Theta\|_\infty \sqrt{\frac{d\log d}{n}}\right)^2 \qquad \text{for all } \Theta \in \mathbb{R}^{d\times d} \tag{78}$$

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

## E.1 Proof of Lemma 8

Applying Proposition 2 to $\widehat{\Delta}^t$ and using the fact that $d\|\widehat{\Delta}^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq |\!|\!|\widehat{\Delta}^t|\!|\!|_F^2 - c_1\alpha|\!|\!|\widehat{\Delta}^t|\!|\!|_1 \sqrt{\frac{d\log d}{n}} - c_2\,\alpha^2\frac{d\log d}{n}, \tag{79}$$

where we recall our convention of allowing the constants to change from line to line. From Lemma 1,

$$|\!|\!|\widehat{\Delta}^t|\!|\!|_1 \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\,|\!|\!|\widehat{\Delta}^t|\!|\!|_F + 2|\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1 + 2|\!|\!|\Delta^*|\!|\!|_1 + \Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F.$$

Since $\rho \leq |\!|\!|\Theta^*|\!|\!|_1$, Lemma 5 implies that $|\!|\!|\Delta^*|\!|\!|_1 \leq 2\Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F + |\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1$, and hence that

$$|\!|\!|\widehat{\Delta}^t|\!|\!|_1 \leq 2\,\Psi(\overline{\mathcal{M}}^\perp)\,|\!|\!|\widehat{\Delta}^t|\!|\!|_F + 4|\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F. \tag{80}$$

Combined with the lower bound, we obtain that $\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n}$ is lower bounded by

$$|\!|\!|\widehat{\Delta}^t|\!|\!|_F^2\left\{1 - \frac{2c_1\,\alpha\Psi(\overline{\mathcal{M}}^\perp)\sqrt{\frac{d\log d}{n}}}{|\!|\!|\widehat{\Delta}^t|\!|\!|_F}\right\} - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4|\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

Consequently, for all iterations such that $|\!|\!|\widehat{\Delta}^t|\!|\!|_F \geq 4c_1\Psi(\overline{\mathcal{M}}^\perp)\sqrt{\frac{d\log d}{n}}$, we have

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}|\!|\!|\widehat{\Delta}^t|\!|\!|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4|\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n}.$$

By subtracting off an additional term, the bound is valid for all $\widehat{\Delta}^t$—viz.

$$\frac{\|\mathfrak{X}_n(\widehat{\Delta}^t)\|_2^2}{n} \geq \frac{1}{2}|\!|\!|\widehat{\Delta}^t|\!|\!|_F^2 - 2c_1\,\alpha\sqrt{\frac{d\log d}{n}}\Big\{4|\!|\!|\Pi_{\mathcal{M}^\perp}(\theta^*)|\!|\!|_1 + 5\Psi(\overline{\mathcal{M}}^\perp)|\!|\!|\Delta^*|\!|\!|_F\Big\} - c_2\,\alpha^2\frac{d\log d}{n} - 16c_1^2\alpha^2\Psi^2(\overline{\mathcal{M}}^\perp)\frac{d\log d}{n}$$

## E.2 Proof of Lemma 9

Applying Proposition 2 to $\Gamma^t$ and using the fact that $d\|\Gamma^t\|_\infty \leq 2\alpha$ yields

$$\frac{\|\mathfrak{X}_n(\Gamma^t)\|_2^2}{n} \leq |\!|\!|\Gamma^t|\!|\!|_F^2 + c_1\alpha|\!|\!|\Gamma^t|\!|\!|_1 \sqrt{\frac{d\log d}{n}} + c_2\,\alpha^2\frac{d\log d}{n}, \tag{81}$$

where we recall our convention of allowing the constants to change from line to line. By triangle inequality, we have $|\!|\!|\Gamma^t|\!|\!|_1 \leq |\!|\!|\Theta^t - \widehat{\Theta}|\!|\!|_1 + |\!|\!|\Theta^{t+1} - \widehat{\Theta}|\!|\!|_1 \;=\; |\!|\!|\widehat{\Delta}^t|\!|\!|_1 + |\!|\!|\widehat{\Delta}^{t+1}|\!|\!|_1$. Equation 80 gives us bounds on $|\!|\!|\widehat{\Delta}^t|\!|\!|_1$ and $|\!|\!|\widehat{\Delta}^{t+1}|\!|\!|_1$. Substituting them into the upper bound (81) yields the claim.

## References

[1] A. Agarwal, S. N. Negahban, and M. J. Wainwright. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. Technical report, UC Berkeley, March 2011. Available at arXiv:1102.4807.

[2] A. A. Amini and M. J. Wainwright. High-dimensional analysis of semdefinite relaxations for sparse principal component analysis. *Annals of Statistics*, 5B:2877–2921, 2009.

[3] 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.

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

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

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

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

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

[9] F. Bunea, A. Tsybakov, and M. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, pages 169–194, 2007.

[10] E. J. Candes, X. Li, Y. Ma, and J. Wright. Robust Principal Component Analysis? Technical report, Stanford, 2009. available at arXiv:0912.3599.

[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. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. Technical report, MIT, June 2009. Available at `arXiv:0906.2220v1`.

[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 $\ell_1$-ball for learning in high dimensions. In *International Conference on Machine Learning (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*, New York, NY, USA, 2009. ACM.

[17] E. T. Hale, Y. Wotao, and Y. Zhang. Fixed-point continuation for $\ell_1$-minimization: Methodology and convergence. *SIAM J. on Optimization*, 19(3):1107–1130, 2008.

[18] D. Hsu, S. M. Kakade, and T. Zhang. Robust Matrix Decomposition with Outliers. Technical report, Rutgers University, 2010. available at arXiv:1011.1518.

[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 *International Conference on Machine Learning*, New York, NY, USA, 2009. ACM.

[21] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes.* Springer-Verlag, New York, NY, 1991.

[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] K. Lounici, M. Pontil, A. B. Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. Technical Report arXiv:0903.1468, ETH Zurich, March 2009.

[24] 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.

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

[26] 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.

[27] S. Negahban and M. J. Wainwright. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. In *Proceedings of the ICML Conference*, Haifa, Israel, June 2010.

[28] S. Negahban and M. J. Wainwright. Restricted strong convexity and (weighted) matrix completion: Optimal bounds with noise. Technical report, UC Berkeley, September 2010. Available at http://arxiv.org/abs/1009.2118.

[29] 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, and presented in part at ICML, Haifa, Israel.

[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. Ngai1 and J.-P.. Penot2. Paraconvex functions and paraconvex sets. *Studia Mathematica*, 184:1–29, 2008.

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

[34] 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.

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

[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. Technical Report arXiv:0912.5338v2, Universite de Paris, January 2010.

[38] N. Srebro, N. Alon, and T. S. Jaakkola. Generalization error bounds for collaborative prediction with low-rank matrices. In *Neural Information Processing Systems (NIPS)*, Vancouver, Canada, December 2005.

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

[40] 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.

[41] S. van de Geer. The deterministic lasso. In *Proc. of Joint Statistical Meeting*, 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. Technical report, University of Texas, Austin, 2010. available at arXiv:1010.4237.

[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.

[46] H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.