Provable Low-Rank Plus Sparse Matrix Recovery Via Nonconvex Regularizers

April Sagan [1] **John E. Mitchell** [1]

Abstract

This paper considers a large class of problems where we seek recover a low rank matrix and/or sparse vector from some set of measurements. While methods based off of convex relaxations suffer from a (possibly large) estimator bias, and other nonconvex methods require the rank or sparsity to be known a priori, we utilize nonconvex regularizers to minimize the rank and l_0 norm without the estimator bias from the convex relaxation. We present a novel analysis of the alternating proximal gradient descent algorithm applied to such problems, and bound the error between the iterates and the ground truth sparse and low rank matrices. The algorithm and error bound can be applied to sparse optimization, matrix completion, and robust principle component analysis as special cases of our results. Computationally, our algorithm obtains state of the art performance in each of these problems for various test cases.

1. Introduction

In order to better understand large data-set and to make inferences about them, it is helpful to understand the underlying patterns in the dataset. Even when the underlying patter is highly nonlinear, the data matrix can be approximated as being low rank, an observation that enables techniques to analyze the data in terms of a low dimensional latent space, such as Principle Component Analysis (PCA), identifying outliers through Robust PCA (RPCA), and accurately inferring data points from very few observations of a data matrix through matrix completion.

Data analysis techniques based upon this low rank property have received much attention in the past decade, with impressive computational results on large matrices and theoretical results guaranteeing the success of RPCA and matrix completion. Many of these results are based off of minimizing the nuclear norm of a matrix (defined as the sum of the singular values) as a surrogate for the rank function, similar to minimizing the l_1 norm to promote sparsity in a vector.

While the convex relaxation is an incredibly useful technique in many applications, minimizing the nuclear norm of a matrix has been shown to introduce a (sometimes very large) estimator bias. Intuitively, we expect to see this bias because if we hope to recover a rank r matrix, we must impose enough weight on the nuclear norm term so that the $(r+1)$th singular value is zero. By the nature of the nuclear norm, this requires also putting weight on minimizing the first r singular values, resulting in a bias towards zero proportional to the spectral norm of the noise added to the true data matrix.

Fortunately, recent work has shown that the estimator bias from convex regularizers can be reduced (or even eliminated, for well conditioned matrices) by using nonconvex regularizers, such as the Schatten-p norm or the minimax concave penalty (MCP). It has been shown that for sparse optimization, the nonconvexity introduced from these regularizers does not create a further burden in the the optimization process – in the right circumstances, the nonconvex problem has just one minimizer. Similar results for rank minimization problems have been previously unavailable, and gap that we have aimed to fill in this paper.

1.1. Summary of Contributions

In this paper, we focus on the nonconvex, unconstrained optimization problem:

$$\min_{L,s} \ \lambda_L \Phi_{\gamma_L}(L) + \lambda_s \phi_{\gamma_s}(s) + \frac{1}{2n}||\mathcal{A}_L(L) + A_S s - b||_2^2 \quad (1)$$

We denote $\phi : \mathbb{R} \to \mathbb{R}_+$ to be a concave function used to promote sparsity in both the singular values of L and individual entries in s. We overload the notation to allow for ϕ_{γ_s} to be a function of a vector $x \in \mathbb{R}^{d_s}$ whose range is $\mathbb{R}_+$, and we denote $\Phi_{\gamma_L} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}_+$ as a surrogate to the rank function:

$$\phi_{\gamma_s}(x) = \sum_{i=1}^{d_s} \phi_{\gamma_s}(x_i), \ \ \Phi_{\gamma_L}(X) = \sum_{i=1}^{\min(d_1,d_2)} \phi_{\gamma_L}(\sigma_i(X)).$$

[1]Department of Mathematical Sciences, Rensselaer Polytechnic Institute, Troy, NY 12180. Correspondence to: April Sagan <aprilsagan1729@gmail.com>.

Preliminary work. Under review by the International Conference on Machine Learning. Do not distribute.

where $\sigma_i(X)$ denotes the ith largest singular value of X. We restrict our focus to nonconvex regularizers that are *amenable regularizers*, as described in (Loh & Wainwright, 2015), and defined below. Some key properties of amenable regularizers are stated in Appendix tk.

Definition 1. *A function $\phi_\gamma(t) : \mathbb{R}_+ \to \mathbb{R}_+$ is amenable if it satisfies the following criteria.*

1. $\phi_\gamma(0) = 0$
2. *ϕ_γ is non decreasing*
3. *For $t > 0$, the function $\frac{\phi_\gamma(t)}{t}$ is non increasing in t.*
4. *the function ϕ_γ is differentiable for all $t \neq 0$ and sub-differentiable at $t = 0$ with $\lim_{t\to 0^+} \phi'_\gamma(t) = 1$.*
5. *The function $\phi_\gamma(t)$ is ν weakly convex. That is, the function $\rho_\nu := \phi_\gamma(t) + \frac{\nu}{2}$ is convex.*

The linear mappings $\mathcal{A}_L : \mathbb{R}^{d_1\times d_2} \to \mathbb{R}^n$ and $A_s \in \mathbb{R}^{d_s\times n}$ serve as the *observation models* of the underlying low rank matrices and sparse vectors. Most commonly, we are interested in the observation model

$$\big[\mathcal{A}(X)\big]_k = X_{i_k,j_k}$$

for $(i_k, j_k) \in \Omega^{obs}$, where $\Omega^{obs} \subseteq \{1, \dots d_1\} \times \{1, \dots d_2\}$ is the set of indices where we have a measurement of the low rank matrix we hope to reconstruct.

We present a very simple alternating algorithm to find a stationary point of (1). Though similar algorithms have been previously studied and shown to converge, we present a novel analysis of this algorithm and show that not only does this algorithm linearly converge, but it converges to the exact low rank matrix L^* and sparse vector s^* when no Gaussian noise is present. Furthermore, when Gaussian noise is present, we obtain an error bound that matches the minmax optimal rate. Our bound greatly improves upon bounds obtained in previous results analysing the convex relaxation and quantify the observation based on computational results that nonconvex regularizers reduce the impact of noise on the quality of the estimator.

1.2. Related Works

The matrix completion problem is as follows: given the values of a low rank matrix for only a sparse set of indices, we seek to determine the rest of the values of the matrix. While the problem of finding the minimum rank matrix that fits the observations is NP hard in general, it has been shown that under some assumptions, the global minimizer to the convex problem

$$\min_{X\in\mathbb{R}^{d_1\times d_2}} ||X||_* \text{ s.t. } X_{ij} = M_{ij}\ \forall (i,j) \in \Omega^{obs}$$

is exactly M, where Ω^{obs} is the set of indices of M we have observed. If M is rank r and $\mu-$incoherent (as defined in section 3.1), then with high probability for a set, Ω^{obs} of n indices chosen uniformly at random, M is the unique minimizer to the convex relaxation so long as $n > C\mu r d_1 \log^{1.2}(d_1)$ for some universal constant C (Candes & Tao, 2010; Candès & Recht, 2012). This condition was later improved to $n > C\mu r d_1 \log(d_1)$ (Recht, 2011).

Using a convex relaxation for Robust PCA has similar results. While PCA is a powerful technique, it has been shown to be less reliable when just a sparse set of data points are grossly corrupted, and so the goal of RPCA is to identify and remove such corruptions by separating the data matrix into the sum of a low rank and sparse matrix.

$$\min_{L,s} ||L||_* + \lambda_0||S||_1 \text{ st } L_{ij} + S_{ij} = M_{ij}\ \forall (i,j) \in \Omega^{obs}$$

The convex relaxation was shown to give the exact solution when every entry of M is observed in (Chandrasekaran et al., 2011), and when only partially observed (under the same assumptions necessary in matrix completion) by (Candès et al., 2011).

In many cases of practical interest, our measurements may have some level of noise in addition to being only partially observed or having some corrupted entries. For matrix completion, we can relax the constraint using a penalty formulation as follows:

$$\min_{X\in\mathbb{R}^{d_1\times d_2}} ||X||_* + \sum_{(i,j)\in\Omega^{obs}} \big(X_{ij} - M_{ij}\big)^2$$

Likewise, RPCA can be formulated as solving:

$$\min_{L,S\in\mathbb{R}^{d_1\times d_2}} ||L||_* + \lambda_0||S||_1 + \sum_{(i,j)\in\Omega^{obs}} \big(L_{ij}+S_{ij}-M_{ij}\big)^2$$

Statistical guarantees on the performance of the first of these estimators are discussed in (Candes & Plan, 2010; Negahban & Wainwright, 2012; 2011), and the later in (Agarwal et al., 2012). Specific bounds and assumptions for these works are discussed in Section 4.

In order to reduce the estimator bias for l_1 minimization, () proposed an iteratively reweighted l_1 norm method to place more weight on minimizing smaller entries, and less on entries further from zero. This idea was generalized to minimizing any Amenable regularizes to promote sparsity. Theoretical results on the subject include algorithmic guarantees similar to the ones presented in this paper (Wang et al., 2014), and a proof that the nonconvex problem has no spurious local minimizers (Loh & Wainwright, 2015; 2017).

The same regularizers could be used as a surrogate to the rank function, as originally proposed by (Mohan & Fazel, 2012). In (Lu et al., 2016; Canyi Lu, 2015), the authors propose a generalization of the singular value thresholding

Table 1. Table of common nonconvex regularizers used for l_0 and rank minimization, along with the associated proximal operator.

	$\phi_\gamma(x)$	$\text{prox}^\tau_{\phi_\gamma}(y)$
l_1	$\lvert x\rvert$	$\text{sign}(y)(\lvert y\rvert - \frac{1}{\tau})_+$
MCP	$\begin{cases} \lvert x\rvert - \frac{x^2}{2\gamma} & \lvert x\rvert \le \gamma \\ \frac{\gamma}{2} & \lvert x\rvert > \gamma \end{cases}$	$\text{sign}(y)\text{min}\Big(\lvert y\rvert, \frac{\gamma\tau}{\gamma\tau-1}(\lvert y\rvert - \frac{1}{\tau})_+\Big)$

algorithm proposed by (), which was later applied to the problem of RPCA in (Chartrand, 2012; Kang et al., 2015). For the problem of matrix completion, the algorithms proposed by (Yao et al., 2019) and (**?**) achieve the fastest computational complexity in other state of the art methods.

Other approaches to low rank optimization rely on, instead of minimizing a surrogate to the rank function, constraining the matrix to be a given rank. This can be done using constrained optimization to optimize over the set of all rank r matrices[], or by using the low-rank factorization of a matrix $X = UV^T$ for $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{d_2 \times r}$. For RPCA, we can use a constraint to ensure the matrix S is sparse [], or we can minimize the l_1 norm.

Previous work on theoretical results pertaining to rank and sparsity methods have consisted of algorithmic guarantees ensuring that we can obtain a matrix sufficiently close to the ground truth low rank and/or sparse matrix for both matrix completion (Jain et al., 2013) and RPCA (Netrapalli et al., 2014; Zhang et al., 2018; Yi et al., 2016). Additionally, both of these problems have been shown to have no spurious local minimizers, and so the ground truth matrices are the only minimizers under some assumptions ().

2. Alternating Proximal Gradient Descent

Many different methods have been shown to be effective when minimizing nonconvex relaxations of the l_0 and rank functions, including iterative reweighted methods, and methods based off of low rank factorization. In this paper, we focus on the most commonly used technique: alternating proximal gradient descent.

The proximal gradient descent method applied to the function $F(x) = \phi(x) + g(x)$ iteratively solve the following problem:

$$\begin{aligned} x^{k+1} &= \underset{x}{\text{argmin}}\ \phi(x) + \frac{1}{2\tau}||x - \big(x^k - \tau\nabla g(x^k)\big)||^2 \\ &:= \text{prox}^\tau_\phi\big(x^k - \tau\nabla g(x^k)\big) \end{aligned}$$

where we defined the *proximal operator* of a function as the minimum of a combination of the function and the distance from a given point. For many functions ϕ that we are interested in, the proximal operator has a closed form solution, which we show in Table 1.

For each of the sparsity promoting regularizers in Table 1, the proximal operator is also dubbed as a *shrinkage operator* or a *thresholding operator* because when the input is less than τ, the output is 0. Otherwise, the input is moved towards zero or, for some nonconvex regularizers, is unchanged. So, we can view the proximal gradient algorithm as iteratively taking a step in the gradient direction of $g(x)$, and then applying the proximal operator to promote sparsity.

When applied to the optimization problem in Equation (1), we have

$$\tilde{L}^{k+1} = L^k - \frac{\tau_L}{n}\mathcal{A}_L^*\big(\mathcal{A}_L(L^k) + A_s s^k - b\big) \tag{2a}$$

$$L^{k+1} = \text{prox}^{\tau_L/\lambda_L}_{\Phi_{\gamma_L}}(\tilde{L}^{k+1}) \tag{2b}$$

$$\tilde{s}^{k+1} = s^k - \frac{\tau_S}{n}A_s^T\big(\mathcal{A}_L(L^{k+1}) + A_s s^k - b\big) \tag{2c}$$

$$s^{k+1} = \text{prox}^{\tau_s/\lambda_s}_{\phi_{\gamma_s}}(\tilde{s}^{k+1}) \tag{2d}$$

To further simplify the problem, the following proposition will allow L subproblem to be solved in each singular value separately.

Proposition 1. *Consider the optimization problem*

$$\min_X \sum_i \phi_\gamma(\sigma_i(X)) + \frac{\tau}{2}||X - Y||_F^2 \tag{3}$$

where ϕ_γ is a ν weakly convex function. If $\nu < \tau$, then equation (3) *is strongly convex and the minimizer X^* has the same singular vectors as Y with singular values given by*

$$\begin{aligned} \sigma_i(X^*) &= \underset{x}{\textit{argmin}}\ \phi_\gamma(x) + \frac{\tau}{2}(x - \sigma_i(Y))^2 \\ &:= \textit{prox}^\tau_{\phi_\gamma}(\sigma_i(Y)) \end{aligned}$$

where $\textit{prox}^\tau_{\phi_\gamma}(\sigma_i(Y))$ is the proximal operator.

Proposition 1 tells us that L^k has the singular vectors of $\tilde{L}^k$ and singular values given by

$$\sigma_i(L^k) = \text{prox}^{\tau_L/\lambda_L}_{\phi_\gamma}(\sigma_i(\tilde{L}^k)).$$

Likewise, the subproblem in S can be solved in each entry of s individually.

The slowest operation in the alternating proximal gradient method is by far the singular value decomposition. However, in practice we can reduce the number of operations by calculating the truncated singular value decomposition only using the first r_0 singular values, where r_0 is an upper bound on the rank, and enforce that the remaining singular

values are zero. Alternatively, we can calculate each singular value in descending order and stop when a singular value falls below λ_L, as all remaining singular values will be set to zero by the proximal operator. So, in the case of RPCA where each entry is observed, each iteration has a computational complexity of $\mathcal{O}(d_1 d_2 r_0)$, which matches other state of the art methods. CITE

Algorithm 1 Alternating Proximal Gradient Descent for Low-Rank Plus Sparse Optimization

for $k = 1, \dots$ **do**
$g^{k+1} = -\tau_L \mathcal{A}_L^*\big(\mathcal{A}_L(U^k \Sigma^k V^k) + A_s s^k - b\big)$
$[U^{k+1}, \tilde{\Sigma}^{k+1}, V^{k+1}] = \text{LRSSVD}(U^k, \Sigma^k, V^k, g^{k+1})$
$\Sigma^{k+1} = \text{prox}\big(\tilde{\Sigma}^{k+1}\big)$
$\tilde{s}^k = s^k - \tau_s A_s^T\big(\mathcal{A}_L(U\Sigma V) + A_s s^k - b\big)$
$s^{k+1} = \text{prox}^{\lambda_s}\big(\tilde{s}^{k+1}\big)$
end for

In the case of matrix completion, however, only a spare set of entries of the low rank matrix are observed, which could be used to increase the efficiency by reducing the amount of computation needed to find the singular value decomposition of L at each iterations. For a low rank matrix with a low rank factorization $U\Sigma V$, we refer to the problem of finding the SVD of the rank r approximation to the matrix $U\Sigma V + g$ for a sparse matrix g as Low Rank plus Sparse SVD (LRSSVD).

The LRSSVD task can be accomplished efficiently using the same methods as if we were to find the SVD of any other matrix, such as the Power Iteration method. Recall that the computaitonal complexity of the Power Iteration is limited by the amount of operations needed to multiply the matrix by a vector. Because the computational complexities of calculating both $u(U\Sigma V^T + Y)$ for $u \in \mathbb{R}^{d_1}$ and $(U\Sigma V^T + Y)v$ for $v \in \mathbb{R}^{d_2}$ are $\mathcal{O}((d_1 + d_2)r_0 + n)$, we can calculate the top r_0 singular values and vectors of $X + Y$ with only $\mathcal{O}\big((d_1 + d_2)r_0^2 + nr_0\big)$ operations.

The other operations in Algorithm 1 take no more time than the LRPSSVD. The gradient in the L direction, g^k, requires calculating $\Sigma_{ii}^k U_i^k (V_i^k)^T$ for each entry in the support of $\mathcal{A}_L$ In the case of matrix completion, is nr operations, which matches the computational complexity per iteration for state of the art matrix completion algorithms. ()

3. Analysis of Proximal Gradient Descent Algorithm

In this section, we present the main result of the paper: a recursive bound on the difference of the iterates of the alternating proximal gradient algorithm and the ground truth low rank matrix and sparse vector. We present the bound for the most general case, and give results on specific problems in the following section.

3.1. Restricted Isometry and Orthogonality Properties

In order to bound the error in the output of our algorithm relative to the underlying ground truth low-rank and sparse matrices L^* and s^*, we must first make a number of assumptions about L^*, s^*, and the observation models $\mathcal{A}_L$ and A_s.

First, we must assure that the low rank matrix L^* can be separated from a sparse matrix – that is, L^* is not sparse itself. Not only is this necessary for low-rank plus sparse decomposition, but for the problem of matrix completion, this assumption is necessary to assure that a sparse set of observations is a good representation of the entire matrix. For example, consider the matrix consisting of zeros in every entry besides one entry, where the value is 1. We must observe every entry in the matrix to assure that we can reconstruct the matrix exactly, due to the fact that we must observe the nonzero entry and every entry in its row and column.

To exclude such ill-posed problems from our analysis, we will assume that L^* is *incoherent*, as defined as follows.

Definition 2. *For a rank r matrix L with singular value decomposition $U\Sigma V^T$, for orthonormal matrices $U \in \mathbb{R}^{d_1 \times r}, V \in \mathbb{R}^{d_2 \times r}$, and diagonal matrix $\Sigma \in \mathbb{R}^{r \times r}$, the tangent space of L is defined as*

$$\mathcal{T} = \left\{ UX + YV \middle| X \in \mathbb{R}^{r \times d_2}, Y \in \mathbb{R}^{d_1 \times r} \right\} \quad (4)$$

Furthermore, we say this tangent space is $\mu-$incoherent if

$$||U_{:i}||_2^2 \leq \frac{\mu r}{d_1},\ ||V_{j:}||_2^2 \leq \frac{\mu r}{d_2}\ \forall i, j \quad (5)$$

We define the projection of a matrix onto the sparse space Ω as

$$\mathcal{P}_\Omega(X) = \begin{cases} X_{ij} & (i,j) \in \Omega \\ 0 & \text{else} \end{cases} \quad (6)$$

and onto the tangent space $\mathcal{T}$ as

$$\mathcal{P}_\mathcal{T}(X) = UU^T X + XVV^T - UU^T XVV^T \quad (7)$$

Next, we we discuss the conditions that the observation models $\mathcal{A}_L$ and A_s must satisfy in order to recover the ground truth low rank and sparse matrix, known as the restricted isometry property. Loosely, the RIP states that for any two vectors in Ω (or matrices $\mathcal{T}$), we can obtain a sufficiently accurate estimate of the distance between the two through the observation model A_s (or $\mathcal{A}_L$).

Definition 3. *The linear mapping A_s satisfies the (α, κ) sparse Restricted Isometry Property if, for any x satisfying $||x||_0 \leq \alpha d_s$*

$$(1-\kappa_S)||x||^2 \leq \tau_s||A_s x||^2 \leq (1+\kappa_S)||A_s x||^2$$

for some constant τ_s. Likewise, the linear mapping $\mathcal{A}$ satisfies the (μ, r, κ) low rank Restricted Isometry Property if for any X in a μ-incoherent rank r tangent space $\mathcal{T}$,

$$(1-\kappa_L)||X||_F^2 \leq \tau_L||\mathcal{A}_L X||_F^2 \leq (1+\kappa_L)||X||_F^2$$

for some constant τ_L.

In some cases, it may be more useful to use the following characterization of the RIP, which bounds the difference between the operator $\tau\mathcal{A}^*\mathcal{A}$ and the identity operator when restricted to sparse vectors or low-rank and incoherent matrices.

Proposition 2. *For a linear mapping $\mathcal{A}$ satisfying the (α, κ) sparse RIP,*

$$||\tau_s \mathcal{P}_\Omega A_s^T A_s \mathcal{P}_\Omega - \mathcal{P}_\Omega||^2 \leq \kappa.$$

Likewise, for any linear mapping satisfying the (μ, r, κ) low rank RIP,

$$||\tau_L \mathcal{P}_\mathcal{T} \mathcal{A}_L^* \mathcal{A} \mathcal{P}_\mathcal{T} - \mathcal{P}_\mathcal{T}||^2 \leq \kappa$$

Finally, we discuss the interplay between the set of sparse matrices and the low rank, incoherent tangent space, and their observation models. We hope to be able to separate the measurement vector b into two parts: one in the span of $\mathcal{A}_L\mathcal{P}_\mathcal{T}$, and one in the span of $\mathcal{A}_s\mathcal{P}_\Omega$. In order for this to be able to be done quickly, we require there to be no non-trivial vectors in the intersection of the two sets, which is equivalent to saying that $||\mathcal{P}_\Omega A_s^T \mathcal{A}_L \mathcal{P}_T|| < 1$. Under some assumptions, this norm is actually close to zero, a concept we refer to as restricted orthogonality, which we define here and verify applies to the problems we are interested in in Section 4.

Definition 4. *Linear maps $\mathcal{A}_L$ and $\mathcal{A}_S$ satisfy the restricted orthogonality property over the sets $\mathcal{T}$ and Ω (respectively) when*

$$||\mathcal{P}_\Omega \mathcal{A}_S^* \mathcal{A}_L \mathcal{P}_T||^2 \leq \kappa_{SL}$$
$$||\mathcal{P}_T \mathcal{A}_L^* \mathcal{A}_S \mathcal{P}_\Omega||^2 \leq \kappa_{LS}$$

3.2. Main Result

Define the difference between the iterates of the alternating proximal gradient descent algorithm and the ground truth low rank matrix and sparse vector at iteration k as $\Delta_L^k = L^* - L^k$ and $\Delta_s^k = s^* - s^k$. Our main result in the most general form gives a bound on the norm of Δ_L^k and Δ_s^k in terms of the differences at the previous iteration, Δ_L^{k-1} and Δ_s^{k-1}.

Theorem 1. *Let L^k and s^k be the sequences generated by Algorithm 1. Assume that*

$$b = \mathcal{A}_L(L^*) + A_s s^* + \mathcal{E} \in \mathbb{R}^n,$$

where $L^* \in \mathbb{R}^{d_1 \times d_2}$ is a rank r and $\mu-$ incoherent matrix, and $s^* \in \mathbb{R}^{d_s}$ is a sparse vector with $supp(s^*) = \Omega$, and the linear mappings $\mathcal{A}_L$ and A_s satisfy the $(2r, 3\mu, \kappa_L)-$ low rank RIP and the $(\alpha, \kappa_s)-$ sparse RIP respectively, and together satisfy the ROP with constant κ_{Ls}. If $\lambda_L \geq ||\mathcal{A}_L^*(\mathcal{E})||_2 + ||\mathcal{A}_L^* A_s \Delta_s^{k-1}||_2$, then

$$||\Delta_L^{k+1}||_F^2 \leq \kappa_L||\Delta_L^k||_F^2 + \kappa_{LS}||\Delta_S^k||_F^2 + ||\mathcal{P}_T \mathcal{A}_L^* \mathcal{E}||_F^2 + \lambda_L r \phi'(\sigma_r(L^*) - 2\lambda_L)$$

Likewise, if $\lambda_S \geq ||A_s^T(\mathcal{E})||_\infty + ||A_s^T \mathcal{A}_L \Delta_L^{k-1}||_\infty$, then

$$||\Delta_S^{k+1}||_F^2 \leq \kappa_S||\Delta_S^k||_F^2 + \kappa_{LS}||\Delta_L^{k+1}||_F^2 + ||\mathcal{P}_\Omega A_s^T \mathcal{E}||_F^2 + \lambda_s \alpha d_s \phi'(s_{min} - 2\lambda_s)$$

For each of these bound, we can think of the first term as the estimation error introduced by the noise, and the second term as the approximation error, which accounts for the bias in the regularizer proportional to the derivative of the regularizer. Previous result for the nuclear norm and l_1 norm give similar bounds, but make the concession that the approximation error is the dominating term. Under some circumstances, that term is equal to zero in our bound.

3.3. Proof of Main Result

We start by presenting the the following two lemmas regarding the proximal operator for the sparse and low rank regularizers.

Lemma 1. *Let L^* be a rank r, μ-incoherence matrix whose singular vectors form the tangent space T, and $\bar{L}$ be defined as*

$$\bar{L} = \underset{L \in \mathbb{R}^{d_1 \times d_2}}{argmin} \ \lambda\Phi(L) + ||L - L^* + \delta||_F^2 \tag{8}$$

Where Φ is an at most $\frac{1}{\lambda}$ weakly convex regualarizer satisfying Assumption 1, and $\delta \in \mathbb{R}^{d_1 \times d_2}$ satisfies $||\delta||_2 \leq \lambda$. Define $\Delta_L = \bar{L} - L^*$. Then,

$$||\Delta_L||_F^2 \leq ||\mathcal{P}_T(\delta)||_F + 2\lambda\sqrt{r}\phi'\big(\sigma_r(L^*)\big) \tag{9}$$

A proofs of Lemmas 1 can be found in Appendix **??**.

In order to utilize the RIP and ROP conditions, we need to verify that L^{k+1} is low rank and incoherent, and that s^{k+1} is sparse, which we will do inductively. Assume that L^k is at most rank r, and that its tangent space is 2μ incoherent. Additionally, assume that $\text{supp}\big(S^k\big) \subseteq \text{supp}\big(S^*\big)$. Clearly, these conditions are met at the first iteration by initializing the algorithm with $L^0 = 0$ and $S^0 = 0$.

At iteration $k+1$,

$$\begin{aligned} L^* - \bar{L}^{k+1} =& L^* - \big(L^k - \mathcal{A}_L^*\big(\mathcal{A}_L(L^k) + \mathcal{A}_S(S^k) - b\big)\big) \\ =& L^* - L^k + \mathcal{A}_L^*\mathcal{A}_L L^k + \mathcal{A}_S^*\mathcal{A}_S S^k \\ & - \mathcal{A}_L^*\big(\mathcal{A}_L(L^*) + \mathcal{A}_S(S^*) + \mathcal{E}\big) \\ =& \big(\Delta_L^k - \mathcal{A}_L^*\mathcal{A}_L\Delta_L^k\big) - \big(\mathcal{A}_L^*\mathcal{A}_S\Delta_S^k\big) - \mathcal{A}_L^*\mathcal{E} \end{aligned}$$

where the first equality comes from the definition of b, and the second inequality substitutes $\Delta_S^k = S^* - S^k$ and $\Delta_L^k = L^* - L^k$. By Lemma 1 (along with the triangle inequality), we have that

$$\begin{aligned} ||\Delta_L^{k+1}||_F^2 \leq& ||\mathcal{P}_T\big(\Delta_L^k - \mathcal{A}_L^*\mathcal{A}_L\Delta_L^k\big)||_F^2 \\ & + ||\mathcal{P}_T\mathcal{A}_L^*\mathcal{A}_S\Delta_S^k||_F^2 + ||\mathcal{P}_T\mathcal{A}_L^*\mathcal{E}||_F^2 \\ & + \lambda_L r\phi'(\sigma_r(L^*) - 2\lambda_L) \end{aligned}$$

Let $\mathcal{T}^k$ denote the union of the tangent space of the rank r approximation L^k and $\mathcal{T}$ so that $\Delta_L^k \in \mathcal{T}^k$. Then,

$$\begin{aligned} & ||\mathcal{P}_T\big(\Delta_L^k - \mathcal{A}_L^*\mathcal{A}_L\Delta_L^k\big)||_F^2 \\ & \leq ||\mathcal{P}_{\mathcal{T}^k}\big(\Delta_L^k - \mathcal{A}_L^*\mathcal{A}_L\Delta_L^k\big)||_F^2 \\ & \leq ||\mathcal{P}_{\mathcal{T}^k}\big(\mathcal{A}_L^*\mathcal{A}_L - \mathcal{I}\big)\mathcal{P}_{\mathcal{T}^k}\Delta_L^k||_F^2 \leq \kappa_L||\Delta_L^k||_F^2 \end{aligned}$$

where the first inequality comes from the contractive property of $\mathcal{P}_{\mathcal{T}}$, and the second inequality comes from the fact that $\mathcal{P}_{\mathcal{T}^k}\Delta_L^k = \Delta_L^k$. And, because $\mathcal{T}^k$ is has incoherence at most 3μ, we can apply the low-rank RIP to obtain the third inequality.

By the inductive hypothesis stating that $\text{supp}\big(S^k\big) \subseteq \text{supp}\big(S^*\big)$ (and thus, that S^k is α-sparse), we can use the ROP to claim that

$$||\mathcal{P}_{\mathcal{T}}\mathcal{A}_L^*\mathcal{A}_S\Delta_S^k||_F^2 \leq \kappa_{SL}||\Delta_S^k||_F^2.$$

Combining these gives the desired bound on Δ_L^k.

Now, we must show that L^{k+1} is rank r and 2μ incoherent. By the tk inequality (Lemma tk in Appendix tk), we know that

$$\begin{aligned} \sigma_{r+1}(\tilde{L}^{k+1}) \leq& \sigma_{r+1}(L^*) + ||L^* - \tilde{L}^{k+1}||_2 \\ \leq& ||\Delta_s^k||_2 + ||\mathcal{A}_L^*\mathcal{E}||_2 \end{aligned}$$

Because this is less than λ_L by our assumptions, L^{k+1} must be rank r.

In order to show that $\tilde{L}^{k+1}$ is at most 2μ incoherent, we apply the following Lemma, which utilizes the David-Kahan inequality.

Lemma 2. *If X is a rank r, $\mu-$ incoherent matrix, and Δ satisfies $||\Delta||_2 \leq \frac{1}{4}\sigma_r(X)$, then the top r singular vectors of the matrix $X + \Delta$ form a $2\mu-$ incoherent tangent space.*

A proof of Lemma 2 can be found in Appendix (tk). We conclude the proof by bounding Δ_s^k and showing $s^{k+1} \in \Omega$ in a similar manner in Appendix tk.

4. Results for Specific Models

In this section, we utilize Theorem 1 to analyze an application of the alternating proximal gradient descent algorithm to the problems of matrix completion and RPCA.

4.1. Matrix Completion

The proximal gradient descent algorithm applied to matrix completion is outlined in Algorithm **??**.

Lemma 3 ((Recht, 2011))**.** *Let Ω be a set of n entries of $\{1,\dots,d_1\} \times \{1,\dots,d_2\}$ drawn independently at random with uniform probability, with $n > 64\mu r(d_1+d_2)\log(d_2)$. Then, with probability at least $1 - 2d_2^{-2}$,*

$$\frac{5}{6}||X||_F^2 \leq \frac{d_1 d_2}{n}||\mathcal{A}_\Omega(X)||^2 \leq \frac{7}{6}||X||_F^2$$

for any rank r, $\mu-$incoherent matrix X.

Theorem 2. *Let $b = \mathcal{A}_\Omega X^* + \mathcal{E}$ for a rank r, $\mu-$ incoherent matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$. Under the same assumptions as Lemma 3, if $\lambda > ||\mathcal{A}_\Omega^*(\mathcal{E})||_2$, then the iterates of Algorithm tk linearly converge to a point $\bar{X}$ satisfying*

$$||X^* - \bar{X}||_F^2 \leq \underbrace{\frac{6}{5}||\mathcal{P}_{\mathcal{T}}\mathcal{A}_\Omega^*\mathcal{E}||_F^2}_{\substack{\text{optimal} \\ \text{error rate}}} + \underbrace{\frac{6}{5}\lambda r\phi'(\sigma_{r+1}(X^*) - 2\lambda)}_{\text{bias term}}$$

with convergence rate $\frac{1}{6}$.

A proof of Theorem 2 can be found in Appendix tk. The two terms of the error bound account for the optimal error rate and a bias term. The optimal error rate is the error bound if we know the tangent space of X^* a priori, that is, the difference between X^* and the solution to the optimization problem

$$\min_{X \in \mathcal{T}} ||\mathcal{A}_\Omega(X) - b||_F^2.$$

The oracle rate is further discussed in (Candes & Plan, 2010) and (Negahban & Wainwright, 2012).

Perhaps counter-intuitively, a choice of step size, τ, that minimizes the loss function (i.e. the steepest descent step size) is not always the step size that leads to the fastest convergence rate. To see this, we compare the error bound at iteration k in both cases. The steepest descent step size ($\tau = n$) would give

$$\begin{aligned} ||L^* - \tilde{L}^k|| &= L^* - L^{k-1} - \mathcal{P}_\Omega(L^* - L^{k-1}) \\ &= ||P_{\Omega^C}(\Delta_L)|| \end{aligned}$$

and the stepsize informed by the RIP ($\tau = \frac{d_1 d_2}{n}$) gives

$$\begin{aligned} ||L^* - \tilde{L}^k|| &= L^* - L^{k-1} - \frac{d_1 d_2}{n}\mathcal{P}_\Omega(L^* - L^{k-1}) \\ &= ||\Delta_L - \frac{d_1 d_2}{n}P_\Omega(\Delta_L)||. \end{aligned}$$

While the $\tau = 1$ gives a significantly smaller error when simply comparing $L^* - \tilde{L}^k$, the error bound for $L^* - L^k$ comes from projecting $L^* - \tilde{L}^k$ onto $\mathcal{T}$. Without further information, the best bound we can get when using $\tau = n$ would be

$$||L^* - L^k||_F^2 \leq 1 - \frac{(1-\kappa)n}{d_1 d_2}||L^* - L^{k-1}||_F^2.$$

This convergence rate approaches 1 asymptotically when we consider the information theoretic minimum number of measurement for large d_1 and d_2. However, when $\tau = \frac{d_1 d_2}{n}$, the convergence rate remains constant:

$$\begin{aligned}||L^* - L^k||_F^2 &\leq ||\Delta_L - \frac{d_1 d_2}{n}\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}(\Delta_L)||_F^2 \\ &\leq \kappa||\Delta_L||_F^2\end{aligned}$$

The first inequality uses the fact that $\Delta_L \in \mathcal{T}$ and the second uses the RIP.

4.2. Robust PCA

Next, we will use Theorem 1 to analyze the APGD algorithm applied to the problem of RPCA. Specifically, we are interested in the special case of (1) where the $A_s = I_n$, and $\mathcal{A}_L = \mathcal{A}_{\Omega^{obs}}$.

In order for RPCA to be possible, we need the nonzero entries of $\mathcal{A}^*_{\Omega^{obs}} s^*$ to be sufficiently well-distributed throughout the rows and columns – if the sparse corruptions affected the same row or column of L^*, then this would also be a low-rank perturbation and thus be impossible to separate from L^* without further information. So, we will assume that $\mathcal{A}^*_L s^*$ is $\alpha-$sparse, defined as follows.

Definition 5. *The matrix S is α-sparse for $0 < \alpha < 1$ if the proportion of nonzero entries in any row or column is less than α. That is,*

$$||S_{i:}||_0 \leq \alpha d_1 \, , ||S_{:j}||_0 \leq \alpha d_2 \; \forall i, j \tag{10}$$

In order to verify that the ROP property holds, we present the following Lemma:

Lemma 4. *Let $\mathcal{T}$ be a rank r, $\mu-$incoherent tangent space, and let Ω be an $\alpha-$sparse subspace. Then,*

$$||\mathcal{P}_{\mathcal{T}}\mathcal{P}_{\Omega}|| \leq 2\alpha\mu r, \; ||\mathcal{P}_{\Omega}\mathcal{P}_{\mathcal{T}}|| \leq 2\alpha\mu r \tag{11}$$

We can now give a bound for the stationary point of the APGD algorithm applied to RPCA.

Theorem 3. *Let $b = \mathcal{A}_{\Omega^{obs}}(L^*) + s^* + \mathcal{E}$ for a rank r, $\mu-$incoherent matrix $L^* \in \mathbb{R}^{d_1 \times d_2}$, and a sparse vector $s^* \in \mathbb{R}^n$ with $||s^*||_\infty \leq 2||L^*||_\infty$. Under the same assumptions on Ω^{obs} as Lemma 3, and assume that $\mathcal{A}^*_\Omega s^*$ is $\alpha-$ sparse and $\alpha\mu r \leq \frac{1}{64}$. If $\lambda_L \geq \frac{1}{6} + ||\mathcal{A}^*_\Omega \mathcal{E}||_2$ and $\lambda_s \geq \frac{\mu r}{d_1} + ||\mathcal{E}||_\infty$, then the iterates of Algorithm tk linearly converge to a point $\bar{L}, \bar{s}$ satisfying*

$$\begin{aligned}||L^* - \bar{L}||_F^2 \leq \\ ||s^* - \bar{s}^*||^2 \leq\end{aligned}$$

with convergence rate $\frac{1}{6}$.

5. Numerical Results

We implemented Algorithm 1 in both Matlab R2020a, for which the code is available at GitHub. We also made the code available as part of the Python package SpaLoR. All results in this section are obtained with the Matlab version in order to accurately compare to other algorithms which are only available in Matlab, and are run on a Windows 10 desktop with an AMD Phenom 3.40 GHz processor and 8 Gb of RAM.

5.1. Matrix Completion

5.1.1. Synthetic Data

5.1.2. Collaborative Filtering

5.2. Robust PCA

We compare our method to several other prominent RPCA methods, including LMaFit (), AltProj (), IALM (), and (). We compared with many other methods included in the LRSLibrary (), however we only include results from the aforementioned algorithms as they gave the most accurate results for matrices a significant amount with Guassian noise, a test case we emphasise in this section.

It is worth noting that, while our algorithm does not require an estimate of the rank of L^* or the sparsity of S^* a priori, BLANK requires both, and LMaFit and AltProj require an estimate of the rank. However, we found that LMaFit and AltProj still perform very well when this estimate is unreliable, as LMaFit includes a rank-estimation scheme and AltProj starts by performing a rank 1 projection, and increases the rank up until the estimate given. We provide all methods with an upper bound on the rank equal to twice the rank of L^* and an upper bound on the number of corrupted entries equal to twice that of S^*.

5.2.1. Background-Foreground Separation

The most commonly utilized test case in the literature for RPCA is the task of separating the background and foreground of a video. In this scenario, each frame of the video is represented as a column vector in M. If the background is static (or, at least, in some way repetitive), then we can expect the background of each frame to be represented as a low rank matrix, and the foreground as a sparse matrix. See () for further details.

Table 2. table

	TPR	TNR	Accuracy	Time
APGD (ours)	0.922	0.8686	0.8954	13.3
RegL1-ALM	0.9535	0.5409	0.7472	247.78
IALM	0.8805	0.5957	0.7381	245.36

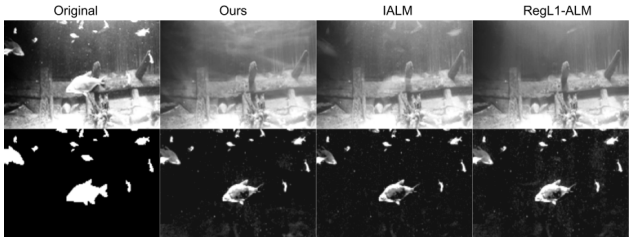

Figure 1. Caption

While surveillance videos from CCTVs () are used as a benchmark in the majority of previous work on RPCA, we evaluated our algorithm on the Underwater Change Detection dataset (). The technical advantages of this test set over the standard benchmarks are two-fold. First, out of the 1100 frames in each video, the ground truth image segmentation is included for last 100 frames. This allows use to present an objective and accurate metric of how well each method is able to identify the forground of the image.

Second, the videos in the Underwater Change Detection dataset present a significantly harder challenge than the relatively clean surveillance footage typically used. In the surveillance video benchmarks, the matrix can be exactly separated into a low rank background matrix and a sparse matrix only including the human subjects, whereas the under water footage includes small moving particulates such as marine snow which is to be filtered out.

6. Conclusions

References

Agarwal, A., Negahban, S., and Wainwright, M. J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. *Ann. Statist.*, 40(2):1171–1197, 04 2012. doi: 10.1214/12-AOS1000. URL https://doi.org/10.1214/12-AOS1000.

Candès, E. and Recht, B. Exact matrix completion via convex optimization. *Commun. ACM*, 55(6):111–119, June 2012. ISSN 0001-0782. doi: 10.1145/2184319.2184343. URL https://doi.org/10.1145/2184319.2184343.

Candes, E. J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010. doi: 10.1109/JPROC.2009.2035722.

Candes, E. J. and Tao, T. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010. doi: 10.1109/TIT.2010.2044061.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *J. ACM*, 58 (3), June 2011. ISSN 0004-5411. doi: 10.1145/1970392.1970395. URL https://doi.org/10.1145/1970392.1970395.

Canyi Lu, Changbo Zhu, C. X. S. Y. Z. L. Generalized singular value thresholding. *Procedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 1805–1811, 2015.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2):572–596, 2011. doi: 10.1137/090761793. URL https://doi.org/10.1137/090761793.

Chartrand, R. Nonconvex splitting for regularized low-rank + sparse decomposition. *IEEE Transactions on Signal Processing*, 60(11):5810–5819, 2012. doi: 10.1109/TSP.2012.2208955.

Jain, P., Netrapalli, P., and Sanghavi, S. Low-rank matrix completion using alternating minimization. In *Proceedings of the Forty-Fifth Annual ACM Symposium on Theory of Computing*, STOC '13, pp. 665–674, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450320290. doi: 10.1145/2488608.2488693. URL https://doi.org/10.1145/2488608.2488693.

Kang, Z., Peng, C., and Cheng, Q. Robust pca via nonconvex rank approximation. In *2015 IEEE International Conference on Data Mining*, pp. 211–220, 2015. doi: 10.1109/ICDM.2015.15.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16(19):559–616, 2015. URL http://jmlr.org/papers/v16/loh15a.html.

Loh, P.-L. and Wainwright, M. J. Support recovery without incoherence: A case for nonconvex regularization. *Ann. Statist.*, 45(6):2455–2482, 12 2017. doi: 10.1214/16-AOS1530. URL https://doi.org/10.1214/16-AOS1530.

Lu, C., Tang, J., Yan, S., and Lin, Z. Nonconvex nonsmooth low rank minimization via iteratively reweighted nuclear norm. *IEEE Transactions on Image Processing*, 25(2): 829–839, 2016. doi: 10.1109/TIP.2015.2511584.

Mohan, K. and Fazel, M. Iterative reweighted algorithms for matrix rank minimization. *J. Mach. Learn. Res.*, 13 (1):3441–3473, November 2012. ISSN 1532-4435.

Negahban, S. and Wainwright, M. J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. *The Annals of Statistics*, pp. 1069–1097, 2011.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *The Journal of Machine Learning Research*, 13(1):1665–1697, 2012.

Netrapalli, P., Niranjan, U., Sanghavi, S., Anandkumar, A., and Jain, P. Non-convex robust pca. In *Advances in Neural Information Processing Systems*, pp. 1107–1115, 2014.

Recht, B. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12(null):3413–3430, December 2011. ISSN 1532-4435.

Wang, Z., Liu, H., and Zhang, T. Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. *Ann. Statist.*, 42(6):2164–2201, 12 2014. doi: 10.1214/14-AOS1238. URL https://doi.org/10.1214/14-AOS1238.

Yao, Q., Kwok, J. T.-Y., and Han, B. Efficient nonconvex regularized tensor completion with structure-aware proximal iterations. volume 97 of *Proceedings of Machine Learning Research*, pp. 7035–7044, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL http://proceedings.mlr.press/v97/yao19a.html.

Yi, X., Park, D., Chen, Y., and Caramanis, C. Fast algorithms for robust pca via gradient descent. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 29, pp. 4152–4160. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/b5f1e8fb36cd7fbeb7988e8639ac79e9-Paper.pdf.

Zhang, X., Wang, L., and Gu, Q. A unified framework for nonconvex low-rank plus sparse matrix recovery. volume 84 of *Proceedings of Machine Learning Research*, pp. 1097–1107, Playa Blanca, Lanzarote, Canary Islands, 09–11 Apr 2018. PMLR. URL http://proceedings.mlr.press/v84/zhang18c.html.

A. proof of theorem 1

Likewise, to bound Δ_S^{k+1},

Lemma 5. *Let S^* be an α-sparse matrix with support Φ, and let $\bar{S}$ be defined as*

$$\bar{S} = \underset{S \in \mathbb{R}^{d_1 \times d_2}}{argmin}\ \lambda\phi(S) + ||S - S^* + \delta||_F^2 \tag{12}$$

Where Φ is an at most $\frac{1}{\lambda}$ weakly convex regualarizer satisfying Assumption 1, and $\delta \in \mathbb{R}^{d_1 \times d_2}$ satisfies $||\delta||_\infty \leq \lambda$. Define $\Delta_S = \bar{S} - S^*$. Then,

$$||\Delta_S||_F^2 \leq ||\mathcal{P}_\Omega(\delta)||_F + 2\lambda\sqrt{|\Omega|}\phi'(S_{min}) \tag{13}$$

$$\begin{aligned} S^* - \bar{S}^{k+1} =& S^* - \big(S^k - \mathcal{A}_S^*\big(\mathcal{A}_S(S^k) + \mathcal{A}_L(L^k) - b\big)\big) \\ =& S^* - S^k + \mathcal{A}_S^*\mathcal{A}_S S^k + \mathcal{A}_L^*\mathcal{A}_L L^k - A_S^*\big(\mathcal{A}_S(S^*) + \mathcal{A}_L(L^*) + \mathcal{E}\big) \\ =& \big(\Delta_S^k - \mathcal{A}_S^*\mathcal{A}_S\Delta_S^k\big) - \big(\mathcal{A}_S^*\mathcal{A}_L\Delta_L^{k+1}\big) - \mathcal{A}_S^*\mathcal{E} \end{aligned}$$

$$||\Delta_S^{k+1}||_F^2 \leq ||\mathcal{P}_\Omega\big(\Delta_S^k - \mathcal{A}_S^*\mathcal{A}_S\Delta_S^k\big)||_F^2 + ||\mathcal{P}_\Omega\mathcal{A}_S^*\mathcal{A}_L\Delta_L^{k+1}||_F^2 + ||\mathcal{P}_\Omega\mathcal{A}_S^*\mathcal{E}||_F^2 + bias\_term$$

$$||\mathcal{P}_\Omega\big(\Delta_S^k - \mathcal{A}_S^*\mathcal{A}_S\Delta_S^k\big)||_F^2 \leq \kappa_S||\Delta_S^k||_F^2$$

$$||\mathcal{P}_\Omega\mathcal{A}_S^*\mathcal{A}_L\Delta_L^{k+1}||_F^2 = ||\mathcal{P}_\Omega\mathcal{A}_S^*\mathcal{A}_L\mathcal{P}_{T^{k+1}}\Delta_L^{k+1}||_F^2 \leq \kappa_{SL}||\Delta_L^{k+1}||_F^2$$

B. Properties of the Proximal Operator

Proof of Proposition 1:

Proof. By convexity of ϕ,

$$\phi(\lambda_i(X_2)) \geq \phi(\lambda_i(X_1)) + \phi'(\lambda_i(X_1))(\lambda_i(X_2) - \lambda_i(X_1))$$

Summing over all $i = 1, \dots n$,

$$\Phi(X_2) \geq \Phi(X_1) + \langle \nabla\phi(\lambda(X_1)), \lambda(X_1) - \lambda(X_2)\rangle$$

By Corollary 1,

$$\Phi(X_2) \geq \Phi(X_1) + \langle U_1\nabla\phi(\lambda(X_1))U_1^T, X_1 - X_2\rangle$$

And, as $\nabla\Phi(X) = U\nabla\phi(\lambda(X))U^T$, Φ is convex.

Consider if the global optimizer X^* was not of the form $U\Sigma V^T$, for a diagonal matrix Σ. Let $\bar{X} = U\Sigma^*V^T$, where $\Sigma_{ii}^* = \sigma_i(X^*)$. By the Hoffman-Wielandt inequality,

$$||X^* - Y||_F^2 \leq \sum_i \big(\sigma_i(X^*) - \sigma_i(Y^*)\big)^2 = ||X^* - Y||^2.$$

And, because $\Phi(\bar{X}) = \Phi(X^*)$, $\bar{X}$ must also be a global minimizer, contradicting the premise that X^* is the sole global minimizer. □

Lemma 6. *Let S^0 be the zero matrix, and assume that $\{\kappa_L^t\}$ is a decreasing series with $\kappa_L^t \leq \frac{\lambda^t}{\beta}$. Then, for all t,*

$$supp(S^t) \subseteq supp(S^*) \tag{14}$$

Furthermore, if $\delta \geq \frac{\lambda^t}{\beta}$ and $\kappa_L^t \leq \delta - \frac{\lambda^t}{\beta}$, then $supp(S^t) = supp(S^*)$.

Proof. We need to show that if $S^*_{ij} = 0$, then $S^{t+1}_{ij} = 0$. By the S^{t+1} update,

$$\begin{aligned}|S^{t+1}_{ij}| =&\big(|M_{ij} - L^{t+1}_{ij}| - \frac{W_0\lambda^t}{\beta^t}\big)_+\\ =&\big(|S^*_{ij} - \Delta^{t+1}_{L_{ij}}| - \frac{W_0\lambda^t}{\beta^t}\big)_+\\ \leq&\big(\kappa^{t+1}_L - \frac{W_0\lambda^t}{\beta^t}\big)_+\end{aligned}$$

By our assumption that $\kappa^{t+1}_L \leq \frac{W_0\lambda^t}{\beta^t}$, $S^{t+1}_{ij} = 0$.

To show the second statement, consider $S^*_{ij} \neq 0$. The S^{t+1} update now gives

$$\begin{aligned}|S^{t+1}_{ij}| =&\big(|S^*_{ij} - \Delta^{t+1}_{L_{ij}}| - \frac{W^t_{ij}\lambda^t}{\beta^t}\big)_+\\ \geq&\big(\delta - \kappa^{t+1}_L - \frac{W_0\lambda^t}{\beta^t}\big)_+\end{aligned}$$

If $\kappa^t_L \leq \delta - \frac{\lambda^t}{\beta}$, then $|S^{t+1}_{ij}| \geq 0$ □

Lemma 7.

$$||\Delta^t_S||_F \leq c\frac{|\Omega|}{mn}||\Delta^t_L||_F + \alpha n\frac{\lambda_s}{\beta}$$

Proof. **case 1** $|S^*_{ij} + \Delta_{L_ij} + E_{ij}| \geq \frac{\lambda_0}{\beta}$ and $(i,j) \in \Omega$
By first order necessary conditions for optimality,

$$\begin{aligned}\bar{S}_ij =&S^*_{ij} + \Delta_{L_ij} + E_{ij} - \rho'(\bar{S}_ij)\\ \leq&S^*_{ij} + \Delta_{L_ij} + E_{ij} - \rho'(S^*_{ij} + \Delta_{L_ij} + E_{ij})\\ \leq&S^*_{ij} + \Delta_{L_ij} + E_{ij} - \rho'(S_{\text{min}} - ||\Delta_L||_\infty - ||E||_\infty)\end{aligned}$$

This gives the bound:

$$|\Delta_{S_{ij}}| \leq |\Delta_{L_ij}| + |E_{ij}| + \rho'(S_{\text{min}} - ||\Delta_L||_\infty - ||E||_\infty)$$

case 2:$|S^*_{ij} + \Delta_{L_ij} + E_{ij}| \geq \frac{\lambda_0}{\beta}$ and $(i,j) \in \Omega$
By assumption that $S_{\text{min}} \geq \frac{\lambda_0}{\beta} + ||\Delta_L||_\infty + ||E||_\infty$, this is a contradiction.

case 3: $S^*_{ij} = 0$
Because we assume $\frac{\lambda_0}{\beta} \geq ||\Delta_L||_\infty + ||E||_\infty$, then $\bar{S}_{ij} = 0$, and $|\Delta_{S_{ij}}| = 0$.

Combining these facts,

$$||\Delta_S||_F \leq ||\mathcal{P}_\Omega(\Delta_L + \frac{\lambda_s}{\beta})||_F \leq ||\mathcal{P}_\Omega(\Delta_L)||_F + |\Omega|\rho'(S_{\text{min}} - ||\Delta_L||_\infty - ||E||_\infty)$$

□

To bound Δ^t_L, we utilize strong convexity of the L^t subproblem.

Lemma 8.

$$||\Delta^t_L||_F \leq \frac{4\mu r}{n}||\Delta_S||_F + \frac{2\sqrt{r}\rho'(\sigma_r(L^*))}{\beta}$$

Proof. Let

$$\begin{aligned} f_L^t =& \rho(L) + \frac{\beta}{2}||L + S^{t-1} - M||_F^2 \\ =& \rho(L) + \frac{\beta}{2}||L - \left(L^* + \Delta_S^{t-1}\right)||_F^2 \end{aligned}$$

be the function minimized to obtain L^t, with $0 \in \partial f_L^t(L^t)$. Every subgradient of f_L^t at L^* is of the form

$$U\Sigma^\rho V^T + W + \beta\Delta_S^{t-1} \in \partial f_L^t(L^*) \forall W \in T^\perp, ||W||_2 \leq 1$$

where $L^* = U\Sigma V^T$ and

$$\Sigma^\rho = \text{diag}(\rho'(\sigma_1(L^*)), \rho'(\sigma_2(L^*)), \ldots, \rho'(\sigma_r(L^*))).$$

By strong convexity, for any W satisfying $W \in T^\perp$ and $||W||_2 \leq 1$, we have that

$$\begin{aligned} ||L^* - L^t||_F \leq& \frac{1}{\beta - \mu}||U\Sigma^\rho V^T + W + \beta\Delta_S^{t-1}||_F \\ \leq& \frac{\sqrt{r}\rho'(\sigma_r(L^*))}{\beta - \mu} + \frac{1}{\beta - \mu}||W + \beta\Delta_S^{t-1}||_F \\ \leq& \frac{\sqrt{r}\rho'(\sigma_r(L^*))}{\beta - \mu} + \frac{1}{\beta - \mu}||\beta\Delta_S^{t-1} - \mathcal{P}_{T^\perp, ||\cdot||\leq 1}(\Delta_S^{t-1})||_F \end{aligned}$$

Because $\beta||\Delta_S||_2 \leq 1$ by assumption,

$$||L^* - L^t||_F \leq \frac{\sqrt{r}\rho'(\sigma_r(L^*))}{\beta - \mu} + \frac{\beta}{\beta - \mu}||\mathcal{P}_T(\Delta_S^{t-1})||_F$$

Consider the case where the unique global optimizer of (**??**) □

Lemma 9. *If $||\Delta_S||_2 \leq \frac{\mu r}{n}\lambda_r(L^*)$, Then L^t is 3μ incoherent.*

Proof. By the update equation for L^t, the eigenvectors of L^t are those of $L^* + \Delta_S$. Let $U \in \mathbb{R}^{n\times r}$ and $\tilde{U} \in \mathbb{R}^{n\times r}$ denote the eigenvectors of L^* and L^t respectively. By the Davis Kahan theorem,

$$\text{dist}\left(U, \tilde{U}\right) \leq \frac{||\Delta_S||_2}{\lambda_r(L^*) + ||\Delta_S||_2} \leq \frac{\mu r}{n}.$$

Let u_i and $\tilde{u}_i$ be the i^th eigenvector of L^* and L^t respectively, and let θ_i be the angle between the vectors.

$$\max(\tilde{u}_i) \leq \max(u_i) + 2\sin(\theta_i) \leq \frac{\mu r}{n} + \frac{2\mu r}{n} = \frac{3\mu r}{n}$$

Therefore, L^t is $\mu + 2$ incoherent. □

Lemma 10. *If $\frac{\beta}{2\gamma_L}||\Delta_S^{k-1} + E||_2 \leq \frac{1}{2}$, then*

$$||\Delta_L^k||_* \leq 4\sqrt{r}||\Delta_L^k||_F \tag{15}$$

Proof. By global optimality of the L subproblem,

$$\rho(L^k) + \frac{\beta}{2}||L^k + S^{k-1} - \tilde{M}||_F^2 \leq \phi(L^*) + \frac{\beta}{2}||L^* - \tilde{M}||_F^2$$

$$\begin{aligned}||L^k + S^{k-1} - \tilde{M}||_F^2 =& ||\Delta_L^k - \Delta_S^{k-1} + E||_F^2 \\ =& ||\Delta_L^k||_F^2 + 2\langle \Delta_L^k, \Delta_S^{k-1} + E\rangle + ||\Delta_S^{k-1} + E||_F^2\end{aligned}$$

$$||L^* - \tilde{M}||_F^2 = ||\Delta_S^{k-1} + E||_F^2$$

$$\begin{aligned}\frac{\beta}{2}||\Delta_L^k||_F^2 &\le \phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k)) + \beta\langle \Delta_L^k, \Delta_S^{k-1} + E\rangle \\ &\le \phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k)) + \frac{\beta}{2}||\Delta_S^{k-1} + E||_2||\Delta_L^k||_* \\ &\le \phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k)) + \frac{\beta}{\gamma L}||\Delta_S^{k-1} + E||_2\big(\phi_\gamma(\Delta_L) + \frac{\mu_L}{2}||\Delta_L||_F^2\big)\end{aligned}$$

$$\begin{aligned}0 \le \frac{\beta - \mu_L}{2} &\le \phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k)) + \frac{1}{2}\phi_\gamma(\Delta_L) \\ &\le \phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k)) + \frac{1}{2}\big(\phi_\gamma(\sigma(L^*)) + \phi_\gamma(\sigma(L^k))\big) \\ &\le \frac{1}{2}\big(3\phi_\gamma(\sigma(L^*)) - \phi_\gamma(\sigma(L^k))\big)\end{aligned}$$

$$||P_r(\Delta_L^k)||_* \le 3||\Delta_L^k - P_r(\Delta_L^k)||_*$$

$$\begin{aligned}||\Delta_L^k||_* &= ||P_r(\Delta_L^k)||_* + ||\Delta_L^k - P_r(\Delta_L^k)||_* &(16)\\ &\le 4||P_r(\Delta_L^k)||_* \le 4\sqrt{r}||P_r(\Delta_L^k)||_F \le 4\sqrt{r}||\Delta_L^k||_F &(17)\end{aligned}$$

□

C. Proofs for Specific Models

Lemma 11. *Let T be the tangent space of a rank r matrix with incoherence μ. If S has no more than α non*

$$||\mathcal{P}_T(S)||_F \le 2\alpha\mu r||S||_F$$

Proof. By the triangle inequality,

$$||P_T(S)||_F^2 \le ||UU^TS||_F^2 + ||SVV^T||_F^2 + ||UU^TSVV^T||_F^2 \tag{18}$$

Because U is an orthonormal matrix,

$$\begin{aligned}||UU^TS||_F^2 =& \text{trace}(UU^TSS^TUU^T) = \text{trace}(U^TUU^TSS^TU) &(19)\\ =& \text{trace}(U^TSS^TU) = ||U^TS||_F^2 &(20)\end{aligned}$$

$$||U^T S||_F^2 = \sum_{k=1}^{r}\sum_{i=1}^{d_1}\langle U_k, S_i\rangle^2 \tag{21}$$

$$\leq \sum_{k=1}^{r}\sum_{i=1}^{d_1}\big(\sum_{j\in\Omega_i} U_{kj}^2\big)||S_i||^2 \tag{22}$$

$$= \sum_{i=1}^{d_1}||S_i||^2\sum_{j\in\Omega_i}||U_j||^2 \tag{23}$$

$$\leq \sum_{i=1}^{d_1}||S_i||^2\alpha\mu r \leq \alpha\mu r||S||_F^2 \tag{24}$$

□

Instead of showing the rank of Δ_L is small, we can use ratio

$$\zeta(X) = \frac{||X||_*}{||X||_F}.$$

If X is rank r, then $\zeta X \leq \sqrt{r}$. Additionally, we define the *relative spikiness ratio* as

$$\eta(X) = \frac{||X||_\infty}{||X||_F}.$$

Lemma 12. *There are universal constants (c_0, c_1, c_2, c_3) such that as long as $|\Omega| \geq c_3 n\log(n)$, we have that for all X such that $\eta(X)\zeta(X) \leq \frac{1}{c_0}\sqrt{\frac{|\Omega|}{n\log n}}$ and $\zeta(X) \leq c_4\sqrt{|\Omega|}$,*

$$\frac{8}{9}||X||_F^2 \leq \frac{||\mathcal{P}_\Omega(X)||_F^2}{p} \leq \frac{10}{9}||X||_F^2 \tag{25}$$

with probability $1 - c_1\exp(-c_2 n\log(n))$.

D. Amenable Regularizers

We record the following statements about amenable regularizers, as stated by Loh and Wainwright. For a vector $x \in \mathbb{R}^p$, $\phi_\gamma(x) = \sum_{i=1}^p \phi_\gamma(\beta_i)$.

Lemma 13 ((?)). *For any function ϕ_γ satisfying Assumption 1,*

$$\gamma L||x||_1 \leq \phi_\gamma(x) + \frac{\nu}{2}||x||_2^2 \tag{26}$$

Lemma 14 ((?)). *Suppose ϕ_γ satisfies Assumption 1. Let $v \in \mathbb{R}^p$, and let A denote the index set of the k largest elements of v in magnitude. Suppose $\xi > 0$ is such that $\xi\phi_\gamma(v_A) - \phi_\gamma(v_{A^c}) \geq 0$. Then,*

$$\xi\phi_\gamma(v_A) - \phi_\gamma(v_{A^c}) \leq \lambda L(\xi||v_A||_1 - ||v_{A^c}||_1). \tag{27}$$

Moreover, if $x^* \in \mathbb{R}^p$ is k-sparse, then, for a vector $x \in \mathbb{R}^p$ such that $\xi\phi_\gamma(x^*) - \phi_\gamma(x) \geq 0$ and $\xi > 1$, we have

$$\xi\phi_\gamma(x^*) - \phi_\gamma(x) \leq \lambda L(\xi||v_A||_1 - ||v_{A^c}||_1). \tag{28}$$

where $v = x - x^*$ and A is the index set of the k largest elements of v in magnitude.

E. Useful facts about Eigenvectors and Eigenvalues

Lemma 15. *The matrix 2-norm and nuclear norm are dual to each other. That is,*

$$\langle X, Y\rangle \leq ||X||_2||Y||_* \tag{29}$$

Furthermore, for every Y there is an X with $||X||_2 = 1$ such that

$$\langle X, Y\rangle = ||Y||_*$$

Isn't this Holder's Inequality? It is also a corollary of the Hoffman-Wielandt Inequality.

Lemma 16 ((**?**))**.** *Let v_i denote the eigenvector corresponding to the i^th eigenvalue of X.*

$$\frac{d}{dX}\lambda_i(X) = v_i v_i^T \tag{30}$$

This property allows us to easily compute the derivative of the objective function.

$$\nabla F(X) = \sum f'(\lambda_i(X))v_i v_i^T = V\text{diag}(f'(\lambda(X)))V^T \tag{31}$$

Lemma 17 ((**?**))**.**

$$\sum_{i=1}^{k}\lambda_i(A-B) \leq \sum_{i=1}^{k}\lambda_i(A) - \lambda_i(B) \quad \forall k = 1, .., n \tag{32}$$

Lemma 18 ((**?**))**.**

$$\sum_{i=1}^{n}|\lambda_i(A-B)| = \sum_{i=1}^{n}|\lambda_i(A) - \lambda_i(B)| \tag{33}$$

Theorem 4 (Hoffman-Wielandt Inequality)**.** *Let A and B both be normal matrices in $\mathbb{C}^{n\times n}$. There exists a permutation $\sigma(\cdot)$ on the integers $1, \dots n$ such that*

$$\sum_{i=1}^{n}|\lambda_{\sigma(i)}(A) - \lambda_i(B)|^2 \leq ||A-B||_F^2 \tag{34}$$

Corollary 1. *If A and B are symmetric, then*

$$||A-B||_F^2 \geq \sum_i(\lambda_i(A) - \lambda_i(B))^2 \tag{35}$$

$$\langle A, B\rangle_F \leq \langle\lambda(A), \lambda(B)\rangle \tag{36}$$

Proof. The first corollary is true because, given real eigenvalues, the optimal permutation is simply $\sigma(i) = i$ for $i = 1, \dots, n$. The second inequality comes from expanding the terms in (35).

$$||A||_F^2 + ||B||_F^2 - 2\langle A, B\rangle \geq ||\lambda(A)||^2 + ||\lambda(B)||^2 - 2\langle\lambda(A), \lambda(B)\rangle \tag{37}$$

Because $||A||_F^2 = ||\lambda(A)||^2$, this implies (36). □

Theorem 5 (Davis Kahan Thoeorem (Davis and Kahan 1970))**.** *Let M be a rank r positive semidefinite matrix, and let $\tilde{M} = M + H$, where H is a perturbation with $||H||_2 \leq \lambda_r(M)$. Let U and $\tilde{U}$ be the leading r eigenvectors of M and $\tilde{M}$ respectively. Then,*

$$dist\big(U, \tilde{U}\big) \leq \frac{||H||}{\lambda_r(M) - ||H||}$$

Theorem 6 (Matrix Bernstein)**.** *Consider a sequence of random matrices $\big\{X_l \in \mathbb{R}^{d_1\times d_2}$ such that $\mathbb{E}\big[X_l\big] = 0$ and $||X_l|| \leq B$ for every l. Define*

$$v = max\bigg\{||\mathbb{E}\bigg[\sum_l X_l^T X_l\bigg]||, ||\mathbb{E}\bigg[\sum_l X_l X_l^T\bigg]||\bigg\}$$

Then, for all $\tau \geq 0$,

$$\mathbb{P}\bigg\{||\sum_l X_l \geq \tau\bigg\} \leq (d_1 + d_2)\exp\frac{-\tau^2/2}{v + B\tau/3}.$$

F. More

Lemma 19 (Spectral Estimation)**.** *Let M be a rank r matrix with eigenvectors $U \in \mathbb{R}^{p\times r}$. Let $\tilde{U} \in \mathbb{R}^{p\times r}$ be the top r eigenvectors of the matrix $\frac{|\Omega|}{p^2}\mathcal{P}_\Omega(M) := \tilde{M}$. If $|\Omega| \gtrsim \mu r p \log(p)$, then with probability $1 - p^{10}$,*

$$\textit{dist}\big(U, \tilde{U}\big) \lesssim \frac{1}{\log(p)} \ll 1$$

To prove Lemma 19, we use the Davis Kahan theorem and the Matrix Bernstein theorem.

Proof. Let $X = \tilde{M} - M$. Note that $\mathbb{E}\big[X_{ij}\big] = 0$ and $||X_{ij}|| \leq \frac{p^2}{|\Omega|}||M||_\infty$. By the incoherence of M, $||M||_\infty \leq \frac{\mu r}{p}$, and so

$$||X_{ij}|| \leq \frac{\mu r p}{|\Omega|}$$

$$||\mathbb{E}\Big[\sum_l X_l^T X_l\Big]|| \leq \frac{p\mu^2 r^2}{|\Omega|}$$

The matrix Bernstein inequality gives

$$||M - \tilde{M}|| \lesssim \frac{1}{\log p}.$$

The Davis Kahan theorem states that

$$\text{dist}\big(U, \tilde{U}\big) \leq \frac{||M - \tilde{M}||}{\lambda_r(M) - ||M - \tilde{M}||} \lesssim \frac{1}{\log(p)} \ll 1$$

□