Guaranteed Rank Minimization via Singular Value Projection

Anonymous Author(s)
Affiliation
Address
email

Abstract

Minimizing the rank of a matrix subject to affine constraints is a fundamental problem with many important applications in machine learning and statistics. In this paper we propose a simple and fast algorithm SVP (Singular Value Projection) for rank minimization under affine constraints (ARMP) and show that SVP recovers the minimum rank solution for affine constraints that satisfy a *restricted isometry property* (RIP). Our method guarantees geometric convergence rate even in the presence of noise and requires strictly weaker assumptions on the RIP constants than the existing methods. We also introduce a Newton-step for our SVP framework to speed-up the convergence with substantial empirical gains. Next, we address a practically important application of ARMP - the problem of low-rank matrix completion, for which the defining affine constraints do not directly obey RIP, hence the guarantees of SVP do not hold. However, we provide partial progress towards a proof of exact recovery for our algorithm by showing a more restricted isometry property and observe empirically that our algorithm recovers low-rank *incoherent* matrices from an almost optimal number of uniformly sampled entries. We also demonstrate empirically that our algorithms outperform existing methods, such as those of [5, 18, 14], for ARMP and the matrix completion problem by an order of magnitude and are also more robust to noise and sampling schemes. In particular, results show that our SVP-Newton method is significantly robust to noise and performs impressively on a more realistic power-law sampling scheme for the matrix completion problem.

1 Introduction

In this paper we study the general affine rank minimization problem (ARMP),

$$\min\ rank(X) \quad s.t \quad \mathcal{A}(X) = b, \quad X \in \mathbb{R}^{m\times n},\ b \in \mathbb{R}^d, \tag{ARMP}$$

where $\mathcal{A}$ is an affine transformation from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^d$.

The affine rank minimization problem above is of considerable practical interest and many important machine learning problems such as matrix completion, low-dimensional metric embedding, low-rank kernel learning can be viewed as instances of the above problem. Unfortunately, ARMP is NP-hard in general and is also NP-hard to approximate ([22]).

Until recently, most known methods for ARMP were heuristic in nature with few known rigorous guarantees. In a recent breakthrough, Recht et al. [24] gave the first nontrivial results for the problem obtaining guaranteed rank minimization for affine transformations $\mathcal{A}$ that satisfy a *restricted isometry property* (RIP). Define the isometry constant of $\mathcal{A}$, δ_k to be the smallest number such that for all $X \in \mathbb{R}^{m\times n}$ of rank at most k,

$$(1-\delta_k)\|X\|_F^2 \le \|\mathcal{A}(X)\|_2^2 \le (1+\delta_k)\|X\|_F^2. \tag{1}$$

The above RIP condition is a direct generalization of the RIP condition used in the compressive sensing context. Moreover, RIP holds for many important practical applications of ARMP such as image compression, linear time-invariant systems. In particular, Recht et al. show that for most natural families of random measurements, RIP is satisfied even for only $O(nk \log n)$ measurements. Also, Recht et al. show that for ARMP with isometry constant $\delta_{5k} < 1/10$, the minimum rank solution can be recovered by the minimum trace-norm solution.

In this paper we propose a simple and efficient algorithm SVP (Singular Value Projection) based on the projected gradient algorithm. We present a simple analysis showing that SVP recovers the minimum rank solution for noisy affine constraints that satisfy RIP and prove the following guarantees. (Independent of our work, Goldfarb and Ma [12] proposed an algorithm similar to SVP. However, their analysis and formulation is different from ours. They also require stronger isometry assumptions, $\delta_{3k} < 1/\sqrt{30}$, than our analysis.)

Theorem 1.1 *Suppose the isometry constant of $\mathcal{A}$ satisfies $\delta_{2k} < 1/3$ and let $b = \mathcal{A}(X^*)$ for a rank-k matrix X^*. Then,* SVP *(Algorithm 1) with step-size $\eta_t = 1/(1 + \delta_{2k})$ converges to X^*. Furthermore,* SVP *outputs a matrix X of rank at most k such that $\|\mathcal{A}(X) - b\|_2^2 \le \epsilon$ and $\|X - X^*\|_F^2 \le \epsilon/(1 - \delta_{2k})$ in at most $\left\lceil \frac{1}{\log((1-\delta_{2k})/2\delta_{2k})} \log \frac{\|b\|^2}{2\epsilon} \right\rceil$ iterations.*

Theorem 1.2 (Main) *Suppose the isometry constant of $\mathcal{A}$ satisfies $\delta_{2k} < 1/3$ and let $b = \mathcal{A}(X^*)+e$ for a rank k matrix X^* and an error vector $e \in \mathbb{R}^d$. Then,* SVP *with step-size $\eta_t = 1/(1 + \delta_{2k})$ outputs a matrix X of rank at most k such that $\|\mathcal{A}(X) - b\|_2^2 \le C\|e\|^2 + \epsilon$ and $\|X - X^*\|_F^2 \le \frac{C\|e\|^2+\epsilon}{1-\delta_{2k}}$, $\epsilon \ge 0$, in at most $\left\lceil \frac{1}{\log(1/D)} \log \frac{\|b\|^2}{2(C\|e\|^2+\epsilon)} \right\rceil$ iterations for universal constants C, D.*

As our SVP algorithm is based on projected gradient descent, it behaves as a first order methods and may require a relatively large number of iterations to achieve high accuracy, even after identifying the correct row and column subspaces. To this end, we introduce a Newton-type step in our framework (SVP-Newton) rather than using a simple gradient-descent step. Guarantees similar to Theorems 1.1, 1.2 follow easily for SVP-Newton using the proofs for SVP. In practice, SVP-Newton performs better than SVP in terms of accuracy and number of iterations.

We next consider an important application of ARMP: the low-rank matrix completion problem (MCP)— given a small number of entries from an unknown low-rank matrix, the task is to complete the missing entries. Note that RIP does not hold directly for this problem. Recently, Candes and Recht [6], Candes and Tao [7] and Keshavan et al. [14] gave the first theoretical guarantees for the problem obtaining exact recovery from an almost optimal number of uniformly sampled entries.

While RIP does not hold for MCP, we show that a similar property holds for *incoherent* matrices [6]. Given our refined RIP and a hypothesis bounding the incoherence of the iterates arising in SVP, an analysis similar to that of Theorem 1.1 immediately implies that SVP optimally solves MCP. We provide strong empirical evidence for our hypothesis and show that that both of our algorithms recover a low-rank matrix from an almost optimal number of uniformly sampled entries.

In summary, our main contributions are:

- Motivated by [11], we propose a projected gradient based algorithm, SVP, for ARMP and show that our method recovers the optimal rank solution when the affine constraints satisfy RIP. To the best of our knowledge, our isometry constant requirements are least stringent: we only require $\delta_{2k} < 1/3$ as opposed to $\delta_{5k} < 1/10$ by Recht et al., $\delta_{3k} < 1/4\sqrt{3}$ by Lee and Bresler [18] and $\delta_{4k} < 0.04$ by Lee and Bresler [17].
- We introduce a Newton-type step in the SVP method which is useful if high precision is critically. SVP-Newton has similar guarantees to that of SVP, is more stable and has better empirical performance in terms of accuracy. For instance, on the Movie-lens dataset [1] and rank $k = 3$, SVP-Newton achieves an RMSE of 0.89, while SVT method [5] achieves an RMSE of 0.98.
- As observed in [23], most trace-norm based methods perform poorly for matrix completion when entries are sampled from more realistic power-law distributions. Our method SVP-Newton is relatively robust to sampling techniques and performs significantly better than the methods of [5, 14, 23] even for power-law distributed samples.
- We show that the affine constraints in the low-rank matrix completion problem satisfy a weaker restricted isometry property and as supported by empirical evidence, conjecture that SVP (as well as SVP-Newton) recovers the underlying matrix from an almost optimal number of uniformly random samples.

- We evaluate our method on a variety of synthetic and real-world datasets and show that our methods consistently outperform, both in accuracy and time, various existing methods [5, 14].

2 Method

In this section, we first introduce our Singular Value Projection (SVP) algorithm for ARMP and present a proof of its optimality for affine constraints satisfying RIP (1). We then specialize our algorithm for the problem of matrix completion and prove a more *restricted isometry property* for the same. Finally, we introduce a Newton-type step in our SVP algorithm and prove its convergence.

2.1 Singular Value Decomposition (SVP)

Consider the following more robust formulation of ARMP (RARMP),

$$\min_X \ \psi(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2 \quad s.t \quad X \in \mathcal{C}(k) = \{X : rank(X) \leq k\}. \tag{RARMP}$$

The hardness of the above problem mainly comes from the non-convexity of the set of low-rank matrices $\mathcal{C}(k)$. However, the Euclidean projection onto $\mathcal{C}(k)$ can be computed efficiently using singular value decomposition (SVD). Our algorithm uses this observation along with the projected gradient method for efficiently minimizing the objective function specified in (RARMP).

Let $\mathcal{P}_k : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ denote the orthogonal projection on to the set $\mathcal{C}(k)$. That is, $\mathcal{P}_k(X) = \operatorname{argmin}_Y\{\|Y - X\|_F : Y \in \mathcal{C}(k)\}$. It is well known that $\mathcal{P}_k(X)$ can be computed efficiently by computing the top k singular values and vectors of X.

In SVP, a candidate solution to ARMP is computed iteratively by starting from the all-zero matrix and adapting the classical projected gradient descent update as follows (note that $\nabla\psi(X) = \mathcal{A}^T(\mathcal{A}(X) - b)$):

$$X^{t+1} \leftarrow \mathcal{P}_k\left(X^t - \eta_t \nabla\psi(X^t)\right) = \mathcal{P}_k\left(X^t - \eta_t \mathcal{A}^T(\mathcal{A}(X^t) - b)\right). \tag{1}$$

Figure 1 presents SVP in more detail. Note that the iterates X^t are always low-rank, facilitating faster computation of the SVD. See Section 3 for a more detailed discussion of computational issues.

Algorithm 1 Singular Value Projection (SVP) Algorithm

Require: $\mathcal{A}, b,$ tolerance ε, η_t for $t = 0, 1, 2, \ldots$
Initialize: $X^0 = 0$ and $t = 0$
repeat
  $Y^{t+1} \leftarrow X^t - \eta_t \mathcal{A}^T(\mathcal{A}(X^t) - b)$
  Compute top k singular vectors of Y^{t+1}: U_k, Σ_k, V_k
  $X^{t+1} \leftarrow U_k \Sigma_k V_k^T$
  $t \leftarrow t + 1$
until $\|\mathcal{A}(X^{t+1}) - b\|_2^2 \leq \varepsilon$

Analysis for Constraints Satisfying RIP

Theorem 1.1 shows that SVP converges to an ϵ-approximate solution of RARMP in $O(\log \frac{\|b\|^2}{\epsilon})$ steps. Theorem 1.2 shows a similar result for the noisy case. The theorems follow from the following lemma that bounds the objective function after each iteration.

Lemma 2.1 *Let X^* be an optimal solution of* (RARMP) *and let X^t be the iterate obtained by SVP at t-th iteration. Then, $\psi(X^{t+1}) \leq \psi(X^*) + \frac{\delta_{2k}}{(1-\delta_{2k})}\|\mathcal{A}(X^* - X^t)\|_2^2$, where δ_{2k} is the rank $2k$ isometry constant of $\mathcal{A}$.*

The lemma follows from elementary linear algebra, optimality of SVD (Eckart-Young theorem) and two simple applications of RIP. We refer to the supplementary material (Appendix A) for a detailed proof. We now prove Theorem 1.1. Theorem 1.2 can also be proved similarly; see supplementary material (Appendix A) for a detailed proof.

Proof of Theorem 1.1 Using Lemma 2.1 and the fact that $\psi(X^*) = 0$, it follows that

$$\psi(X^{t+1}) \leq \frac{\delta_{2k}}{(1-\delta_{2k})}\|\mathcal{A}(X^* - X^t)\|_2^2 = \frac{2\delta_{2k}}{(1-\delta_{2k})}\psi(X^t).$$

Also, note that for $\delta_{2k} < 1/3$, $\frac{2\delta_{2k}}{(1-\delta_{2k})} < 1$. Hence, $\psi(X^\tau) \leq \epsilon$ where $\tau = \left\lceil \frac{1}{\log((1-\delta_{2k})/2\delta_{2k})} \log\frac{\psi(X^0)}{\epsilon} \right\rceil$. Further, using RIP for the rank at most $2k$ matrix $X^\tau - X^*$ we get: $\|X^\tau - X^*\| \le \psi(X^\tau)/(1-\delta_{2k}) \le \epsilon/(1-\delta_{2k})$. Now, the SVP algorithm is initialized using $X^0 = 0$, i.e., $\psi(X^0) = \frac{\|b\|^2}{2}$. Hence, $\tau = \left\lceil \frac{1}{\log((1-\delta_{2k})/2\delta_{2k})} \log \frac{\|b\|^2}{2\epsilon} \right\rceil$.

2.2 Matrix Completion

We first describe the low-rank matrix completion problem formally. For $\Omega \subseteq [m] \times [n]$, let $\mathcal{P}_\Omega : \mathbb{R}^{m\times n} \to \mathbb{R}^{m\times n}$ denote the projection onto the index set Ω. That is, $(\mathcal{P}_\Omega(X))_{ij} = X_{ij}$ for $(i,j) \in \Omega$ and $(\mathcal{P}_\Omega(X))_{ij} = 0$ otherwise. Then, the low-rank matrix completion problem (MCP) can be formulated as follows,

$$\min_X \ \mathrm{rank}(X) \quad s.t \quad \mathcal{P}_\Omega(X) = \mathcal{P}_\Omega(X^*),\ X \in \mathbb{R}^{m\times n}. \tag{MCP}$$

Observe that MCP is a special case of ARMP, so we can apply SVP for matrix completion. We use step-size $\eta_t = 1/(1+\delta)p$, where p is the density of sampled entries and δ is a parameter which we will explain later in this section. Using the given step-size and update (1), we get the following update for matrix-completion:

$$X^{t+1} \leftarrow \mathcal{P}_k \left(X^t - \frac{1}{(1+\delta)p}(\mathcal{P}_\Omega(X^t) - \mathcal{P}_\Omega(X^*)) \right). \tag{2}$$

Although matrix completion is a special case of ARMP, the affine constraints that define MCP, $\mathcal{P}_\Omega$, do not satisfy RIP in general. Thus Theorems 1.1, 1.2 above and the results of Recht et al. [24] do not directly apply to MCP. However, we show that the matrix completion affine constraints satisfy RIP for low-rank *incoherent* matrices.

Definition 2.1 (Incoherence) *A matrix $X \in \mathbb{R}^{m\times n}$ with singular value decomposition $X = U\Sigma V^T$ is μ-incoherent if* $\max_{i,j} |U_{ij}| \le \frac{\sqrt{\mu}}{\sqrt{m}}, \quad \max_{i,j} |V_{ij}| \le \frac{\sqrt{\mu}}{\sqrt{n}}$.

The above notion of incoherence is similar to that introduced by Candes and Recht [6] and also used by [7, 14]. Intuitively, high incoherence (i.e., μ is small) implies that the non-zero entries of X are not concentrated in a small number of entries. Hence, a random sampling of the matrix should provide enough global information to satisfy RIP.

Using the above definition, we prove the following refined restricted isometry property.

Theorem 2.2 *There exists a constant $C \ge 0$ such that the following holds for all $0 < \delta < 1$, $\mu \ge 1$, $n \ge m \ge 3$: For $\Omega \subseteq [m] \times [n]$ chosen according to the Bernoulli model with density $p \ge C\mu^2 k^2 \log n/\delta^2 m$, with probability at least $1 - \exp(-n \log n)$, the following restricted isometry property holds for all μ-incoherent matrices X of rank at most k:*

$$(1-\delta)p\,\|X\|_F^2 \ \le\ \|\mathcal{P}_\Omega(X)\|_F^2 \ \le\ (1+\delta)p\,\|X\|_F^2. \tag{3}$$

Roughly, our proof combines a Chernoff bound estimate for $\|\mathcal{P}_\Omega(X)\|_F^2$ with a union bound over low-rank incoherent matrices. A proof sketch is presented in Section 2.2.1.

Given the above refined RIP, if the iterates arising in SVP are shown to be incoherent, the arguments of Theorem 1.1 can be used to show that SVP achieves exact recovery for low-rank incoherent matrices from uniformly sampled entries. As supported by empirical evidence, we hypothesize that the iterates X^t arising in SVP remain incoherent when the underlying matrix X^* is incoherent.

Figure 1 (d) plots the maximum incoherence $\max_t \mu(X^t) = \sqrt{n}\max_{t,i,j} |U^t_{ij}|$, where U^t are the left singular vectors of the intermediate iterates X^t computed by SVP. The figure clearly shows that the incoherence $\mu(X^t)$ of the iterates is bounded by a constant independent of the matrix size n and density p throughout the execution of SVP. Figure 2 (c) plots the threshold sampling density p beyond which matrix completion for randomly generated matrices is solved exactly by SVP for fixed k and varying matrix sizes n. Note that the density threshold matches the optimal information-theoretic bound [14] of $\Theta(k \log n/n)$.

Motivated by Theorem 2.2 and supported by empirical evidence (Figures 2 (c), (d)) we hypothesize that SVP achieves exact recovery from an almost optimal number of samples for incoherent matrices.

Conjecture 2.3 *Fix μ, k and $\delta \le 1/3$. Then, there exists a constant C such that for a μ-incoherent matrix X^* of rank at most k and Ω sampled from the Bernoulli model with density $p = \Omega_{\mu,k}((\log n)/m)$,* SVP *with step-size $\eta_t = 1/(1+\delta)p$ converges to X^* with high probability. Moreover,* SVP *outputs a matrix X of rank at most k such that $\|\mathcal{P}_\Omega(X) - \mathcal{P}_\Omega(X^*)\|_F^2 \le \epsilon$ after $O_{\mu,k}\left(\left\lceil \log\left(\frac{1}{\epsilon}\right)\right\rceil\right)$ iterations.*

2.2.1 RIP for Matrix Completion on Incoherent Matrices

We now prove the restricted isometry property of Theorem 2.2 for the affine constraints that result from the projection operator $\mathcal{P}_\Omega$. To prove Theorem 2.2 we first show the theorem for a *discrete* collection of matrices using Chernoff type large-deviation bounds and use standard quantization arguments to generalize to the continuous case. We first introduce some notation and provide useful lemmas for our main proof[1]. First, we introduce the notion of α-regularity.

Definition 2.2 *A matrix $X \in \mathbb{R}^{m\times n}$ is α-regular if $\max_{i,j} |X_{ij}| \le \frac{\alpha}{\sqrt{mn}} \cdot \|X\|_F$.*

Lemma 2.4 below relates the notion of regularity to incoherence and Lemma 2.5 proves (3) for a *fixed* regular matrix when the samples Ω are selected independently.

Lemma 2.4 *Let $X \in \mathbb{R}^{m\times n}$ be a μ-incoherent matrix of rank at most k. Then X is $\mu\sqrt{k}$-regular.*

Lemma 2.5 *Fix a α-regular $X \in \mathbb{R}^{m\times n}$ and $0 < \delta < 1$. Then, for $\Omega \subseteq [m]\times[n]$ chosen according to the Bernoulli model, with each pair $(i,j) \in \Omega$ chosen independently with probability p,*

$$\Pr\left[\,\left|\|\mathcal{P}_\Omega(X)\|_F^2 - p\|X\|_F^2\right| \ge \delta p\|X\|_F^2\,\right] \le 2\exp\left(-\frac{\delta^2 pmn}{3\,\alpha^2}\right).$$

While the above lemma shows Equation (3) for a fixed rank k, μ-incoherent X (i.e., $(\mu\sqrt{k})$-regular X using Lemma 2.4), we need to show Equation (3) for *all* such rank k incoherent matrices. To handle this problem, we discretize the space of low-rank incoherent matrices so as to be able to use the above lemma and a union bound. We now show the existence of a small set of matrices $S(\mu,\epsilon) \subseteq \mathbb{R}^{m\times n}$ such that every low-rank μ-incoherent matrix is close to an appropriately regular matrix from the set $S(\mu,\epsilon)$.

Lemma 2.6 *For all $0 < \epsilon < 1/2$, $\mu \ge 1$, $m,n \ge 3$ and $k \ge 1$, there exists a set $S(\mu,\epsilon) \subseteq \mathbb{R}^{m\times n}$ with $|S(\mu,\epsilon)| \le (mnk/\epsilon)^{3\,(m+n)k}$ such that the following holds. For any μ-incoherent $X \in \mathbb{R}^{m\times n}$ of rank k with $\|X\|_2 = 1$, there exists $Y \in S(\mu,\epsilon)$ s.t. $\|Y - X\|_F < \epsilon$ and Y is $(4\mu\sqrt{k})$-regular.*

We now prove Theorem 2.2 by combining Lemmas 2.5, 2.6 and applying a union bound. We present a sketch of the proof but defer the details to the supplementary material (Appendix B).

Proof Sketch of Theorem 2.2 Let $S'(\mu,\epsilon) = \{Y : Y \in S(\mu,\epsilon), Y \text{ is } 4\mu\sqrt{k}\text{-regular}\}$, where $S(\mu,\epsilon)$ is as in Lemma 2.6 for $\epsilon = \delta/9mnk$. Let $m \le n$. Then, by Lemma 2.5 and union bound, for any $Y \in S'(\mu,\epsilon)$,

$$\Pr\left[\left|\|\mathcal{P}_\Omega(Y)\|_F^2 - p\|Y\|_F^2\right| \ge \delta p\|Y\|_F^2\right] \le 2(mnk/\epsilon)^{3(m+n)k}\exp\left(\frac{-\delta^2 pmn}{16\mu^2 k}\right) \le \exp(C_1 nk\log n)\cdot\exp\left(\frac{-\delta^2 pmn}{16\mu^2 k}\right),$$

where $C_1 \ge 0$ is a constant independent of m,n,k. Thus, if $p > C\mu^2k^2\log n/\delta^2 m$, where $C = 16(C_1+1)$, with probability at least $1-\exp(-n\log n)$, the following holds

$$\forall Y \in S'(\mu,\epsilon), \quad \left|\|\mathcal{P}_\Omega(Y)\|_F^2 - p\|Y\|_F^2\right| \le \delta p\|Y\|_F^2. \tag{4}$$

As the statement of the theorem is invariant under scaling, it is enough to show the statement for all μ-incoherent matrices X of rank at most k and $\|X\|_2 = 1$. Fix such a X and suppose that (4) holds. Now, by Lemma 2.6 there exists $Y \in S'(\mu,\epsilon)$ such that $\|Y - X\|_F \le \epsilon$. Moreover,

$$\|Y\|_F^2 \le (\|X\|_F + \epsilon)^2 \le \|X\|_F^2 + 2\epsilon\|X\|_F + \epsilon^2 \le \|X\|_F^2 + 3\epsilon k.$$

Proceeding similarly, we can show that

$$\left|\|X\|_F^2 - \|Y\|_F^2\right| \le 3\epsilon k, \qquad \left|\|\mathcal{P}_\Omega(Y)\|_F^2 - \|\mathcal{P}_\Omega(X)\|_F^2\right| \le 3\epsilon k. \tag{5}$$

Combining inequalities (4), (5) above, with probability at least $1-\exp(-n\log n)$ we have,

$$\left|\|\mathcal{P}_\Omega(X)\|_F^2 - p\|X\|_F^2\right| \le \left|\|\mathcal{P}_\Omega(X)\|_F^2 - \|\mathcal{P}_\Omega(Y)\|_F^2\right| + p\left|\|X\|_F^2 - \|Y\|_F^2\right| + \left|\|\mathcal{P}_\Omega(Y)\|_F^2 - p\|Y\|_F^2\right| \le 2\delta p\|X\|_F^2.$$

The theorem follows using the above inequality.

2.3 SVP-Newton

In this section we introduce a Newton-type step in our SVP method to speed up its convergence. Recall that each iteration of SVP (Equation (1)) takes a step along the gradient of the objective function and then projects the iterate to the set of low rank matrices using SVD. Now, the top k

[1] Detailed proofs of all the lemmas in this section are provided in Appendix B of the supplementary material.

singular vectors (U_k, V_k) of $Y^{t+1} = X^t - \eta_t \mathcal{A}^T(\mathcal{A}(X^t) - b)$ determine the range-space and column-space of the next iterate in SVP. Then, Σ_k is given by $\Sigma_k = Diag(U_k^T(X^t - \eta_t \mathcal{A}^T(\mathcal{A}(X^t) - b))V_k)$. Hence, Σ_k can be seen as a product of gradient-descent step for a quadratic objective function, i.e., $\Sigma_k = \operatorname{argmin}_S \psi(U_k S V_k^T)$. This leads us to the following variant of SVP we call SVP-Newton:[2]

$$\text{Compute top } k\text{-singular vectors } U_k, V_k \text{ of } Y^{t+1} = X^t - \eta_t \mathcal{A}^T(\mathcal{A}(X^t) - b)$$

$$X^{t+1} = U_k \Sigma_k V_k, \quad \Sigma_k = \operatorname*{argmin}_S \Psi(U_k S V_k^T) = \operatorname*{argmin}_S \|\mathcal{A}(U_k \Sigma_k V_k^T) - b\|^2.$$

Note that as $\mathcal{A}$ is an affine transformation, Σ_k can be computed by solving a least squares problem on $k \times k$ variables. Also, for a single iteration, given the same starting point, SVP-Newton decreases the objective function more than SVP. This observation along with straightforward modifications of the proofs of Theorems 1.1, 1.2 show that similar guarantees hold for SVP-Newton as well[3].

Note that the least squares problem for computing Σ_k has k^2 variables. This makes SVP-Newton computationally expensive for problems with large rank, particularly for situations with a large number of constraints as is the case for matrix completion. To overcome this issue, we also consider the alternative where we restrict Σ_k to be a diagonal matrix, leading to the update

$$\Sigma_k = \operatorname*{argmin}_{S, s.t., S_{ij}=0 \text{ for } i \neq j} \|\mathcal{A}(U_k S V_k^T) - b\|^2 \tag{6}$$

We call the above method SVP-NewtonD (for SVP-Newton Diagonal). As for SVP-Newton, guarantees similar to SVP follow for SVP-NewtonD by observing that for each iteration, SVP-NewtonD decreases the objective function more than SVP.

3 Related Work and Computational Issues

The general rank minimization problem with affine constraints is NP-hard and is also NP-hard to approximate [22]. Most methods for ARMP either relax the rank constraint to a convex function such as the trace-norm [8], [9], or assume a factorization and optimize the resulting non-convex problem by alternating minimization [4, 3, 15].

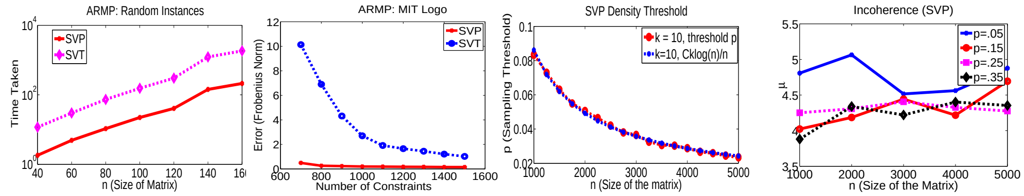
The results of Recht et al. [24] were later extended to noisy measurements and isometry constants up to $\delta_{3k} < 1/4\sqrt{3}$ by Fazel et al. [10] and Lee and Bresler [18]. However, even the best existing optimization algorithms for the trace-norm relaxation are relatively inefficient in practice. Recently, Lee and Bresler [17] proposed an algorithm (ADMiRA) motivated by the *orthogonal matching pursuit* line of work in compressed sensing and show that for affine constraints with isometry constant $\delta_{4k} \leq 0.04$, their algorithm recovers the optimal solution. However, their method is not very efficient for large datasets and when the rank of the optimal solution is relatively large.

For the matrix-completion problem until the recent works of [6], [7] and [14], there were few methods with rigorous guarantees. The alternating least squares minimization heuristic and its variants [3, 15] perform the best in practice, but are notoriously hard to analyze. Candes and Recht [6], Candes and Tao [7] show that if X^* is μ-incoherent and the known entries are sampled uniformly at random with $|\Omega| \geq C(\mu)\, k^2 n \log^2 n$, finding the minimum trace-norm solution recovers the minimum rank solution. Keshavan et.al obtained similar results independently for exact recovery from uniformly sampled Ω with $|\Omega| \geq C(\mu, k)\, n \log n$.

Minimizing the trace-norm of a matrix subject to affine constraints can be cast as a semi-definite program (SDP). However, algorithms for semi-definite programming, as used by most methods for minimizing trace-norm, are prohibitively expensive even for moderately large datasets. Recently, a variety of methods based mostly on iterative soft-thresholding have been proposed to solve the trace-norm minimization problem more efficiently. For instance, Cai et al. [5] proposed a Singular Value Thresholding (SVT) algorithm which is based on Uzawa's algorithm [2]. A related approach based on linearized Bregman iterations was proposed by Ma et al. [20], Toh and Yun [25], while Ji and Ye [13] use Nesterov's gradient descent methods for optimizing the trace-norm.

[2] We call our method SVP-Newton as the Newton method when applied to a quadratic objective function leads to the exact solution by solving the resulting least squares problem.

[3] As a side note, we can show a stronger result for SVP-Newton when applied to the special case of compressed-sensing, i.e., when the matrix X is restricted to be diagonal. Specifically, we can show that under certain assumptions SVP-Newton converges to the optimal solution in $O(\log k)$, improving upon the result of Maleki [21]. We give the precise statement of the theorem and proof in the supplementary material.



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

Figure 1: **(a)** Time taken by SVP and SVT for random instances of the Affine Rank Minimization Problem (ARMP) with optimal rank $k = 5$. **(b)** Reconstruction error for the MIT logo. **(c)** Empirical estimates of the sampling density threshold required for exact matrix completion by SVP (here $C = 1.28$). Note that the empirical bounds match the information theoretically optimal bound $\Theta(k \log n/n)$. **(d)** Maximum incoherence $\max_t \mu(X^t)$ over the iterates of SVP for varying densities p and sizes n. Note that the incoherence is bounded by a constant, supporting Conjecture 2.3.

While the soft-thresholding based methods for trace-norm minimization are significantly faster than SDP based approaches, they suffer from slow convergence (see Figure 2 (d)). Also, noisy measurements pose considerable computational challenges for trace-norm optimization as the rank of the intermediate iterates can become very large (see Figure 3(b)).

For the case of matrix completion, SVP has an important property facilitating fast computation of the main update in equation (2); each iteration of SVP involves computing the singular value decomposition (SVD) of the matrix $Y = X^t + \mathcal{P}_\Omega(X^t - X^*)$, where X^t is a matrix of rank at most k whose SVD is known and $\mathcal{P}_\Omega(X^t - X^*)$ is a sparse matrix. Thus, matrix-vector products of the form Yv can be computed in time $O((m+n)k + |\Omega|)$. This facilitates the use of fast SVD computing packages such as PROPACK [16] and ARPACK [19] that only require subroutines for computing matrix-vector products.

4 Experimental Results

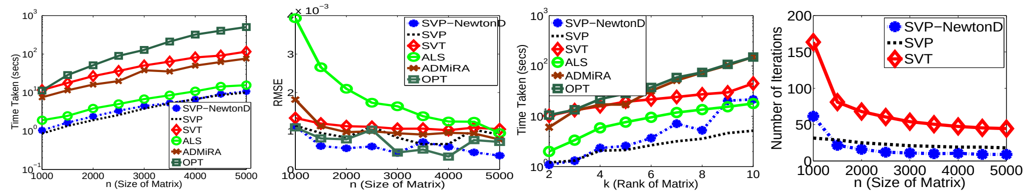
In this section, we empirically evaluate our methods for the affine rank minimization problem and low-rank matrix completion. For both problems we present empirical results on synthetic as well as real-world datasets. For ARMP we compare our method against the trace-norm based singular value thresholding (SVT) method [5]. Note that although Cai et al. present the SVT algorithm in the context of MCP, it can be easily adapted for ARMP. For MCP we compare against SVT, ADMiRA [17], the OptSpace (OPT) method of Keshavan et al. [14], and regularized alternating least squares minimization (ALS). We use our own implementation of SVT for ARMP and ALS, while for matrix completion we use the code provided by the respective authors for SVT, ADMiRA and OPT. We report results averaged over 20 runs. All the methods are implemented in Matlab and use mex files.

4.1 Affine Rank Minimization

We first compare our method against SVT on random instances of ARMP. We generate random matrices $X \in \mathbb{R}^{n\times n}$ of different sizes n and fixed rank $k = 5$. We then generate $d = 6kn$ random affine constraint matrices A_i and compute $b = \mathcal{A}(X)$. Figure 1(a) compares the computational time required by SVP and SVT (in $\log$-scale) for achieving a relative error ($\|\mathcal{A}(X) - \boldsymbol{b}\|_2/\|\boldsymbol{b}\|_2$) of 10^{-3}, and shows that our method requires many fewer iterations and is significantly faster than SVT.

Next we evaluate our method for the problem of matrix reconstruction from random measurements. As in Recht et al. [24], we use the MIT logo as the test image for reconstruction. The MIT logo we use is a 38×73 image and has rank four. For reconstruction, we generate random measurement matrices A_i and measure $b_i = Tr(A_i X)$. We let both SVP and SVT converge and then compute the reconstruction error for the original image. Figure 1 (b) shows that our method incurs significantly smaller reconstruction error than SVT for the same number of measurements.

Matrix Completion: Synthetic Datasets (Uniform Sampling)
We now evaluate our method against various matrix completion methods for random low-rank matrices and uniform samples. We generate a random rank k matrix $X \in \mathbb{R}^{n\times n}$ and generate random Bernoulli samples with probability p. Figure 2 (a) compares the time required by various methods (in $\log$-scale) to obtain a root mean square error (RMSE) of 10^{-3} on the sampled entries for fixed $k = 2$. Clearly, SVP is substantially faster than the other methods. Next, we evaluate our method for increasing k. Figure 2 (b) compares the overall RMSE obtained by various methods. Note that SVP-Newton is significantly more accurate than both SVP and SVT. Figure 2 (c) compares the time required by various methods to obtain a root mean square error (RMSE) of 10^{-3} on the sampled

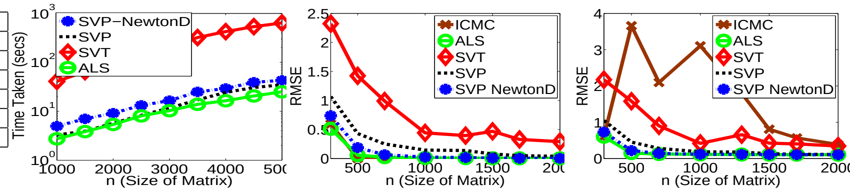


(a) (b) (c) (d)

Figure 2: **(a), (b)** Running time (*on log scale*) and RMSE of various methods for matrix completion problem with sampling density $p = .1$ and optimal rank $k = 2$. **(c)** Running time (*on log scale*) of various methods for matrix completion with sampling density $p = .1$ and $n = 1000$. **(d)** Number of iterations needed to get RMSE 0.001.

k	SVP-NewtonD	SVP	ALS	SVT
2	0.90	1.15	0.88	1.06
3	0.89	1.14	0.87	0.98
5	0.89	1.09	0.86	0.95
7	0.89	1.08	0.86	0.93
10	0.90	1.07	0.87	0.91
12	0.91	1.08	0.88	0.90

(a)



(b) (c) (d)

Figure 3: **(a)**: RMSE incurred by various methods for matrix completion with different rank (k) solutions on Movie-Lens Dataset. **(b)**: Time(*on log scale*) required by various methods for matrix completion with $p = .1$, $k = 2$ and 10% Gaussian noise. Note that all the four methods achieve similar RMSE. **(c)**: RMSE incurred by various methods for matrix completion with $p = 0.1$, $k = 10$ when the sampling distribution follows Power-law distribution (Chung-Lu-Vu Model). **(d)**: RMSE incurred for the same problem setting as plot (c) but with added Gaussian noise.

entries for fixed $n = 1000$ and increasing k. Note that our algorithms scale well with increasing k and are faster than other methods. Next, we analyze reasons for better performance of our methods. To this end, we plot the number of iterations required by our methods as compared to SVT (Figure 2 (d)). Note that even though each iteration of SVT is almost as expensive as our methods', our methods converge in significantly fewer iterations.

Finally, we study the behavior of our method in presence of noise. For this experiment, we generate random matrices of different size and add approximately 10% Gaussian noise. Figure 2 (c) plots time required by various methods as n increases from 1000 to 5000. Note that SVT is particularly sensitive to noise. One of the reason for this is that due to noise, the rank of the intermediate iterates arising in SVT can be fairly large.

Matrix Completion: Synthetic Dataset (Power-law Sampling) We now evaluate our methods against existing matrix-completion methods under more realistic power-law distributed samples. As before, we generate a random rank-$k = 10$ matrix $X \in \mathbb{R}^{n \times n}$ and sample the entries of X using a graph generated using Chung-Lu-Vu model with power-law distributed degrees (see [23]) for details. Figure 3 (c) plots the RMSE obtained by various methods for varying n and fixed sampling density $p = 0.1$. Note that SVP-NewtonD performs significantly better than SVT as well as SVP. Figure 3 (d) plots the RMSE obtained by various methods when each sampled entry is corrupted with around 1% Gaussian noise. Note that here again SVP-NewtonD performs similar to ALS and is significantly better than the other methods including the ICMC method [23] which is specially designed for power-law sampling but is quite sensitive to noise.

Matrix Completion: Movie-Lens Dataset
Finally, we evaluate our method on the Movie-Lens dataset [1], which contains 1 million ratings for 3900 movies by 6040 users. Figure 3 (a) shows the RMSE obtained by each method with varying k. For SVP and SVP-Newton, we fix step size to be $\eta = 1/p\sqrt{(t)}$, where t is the number of iterations. For SVT, we fix $\delta = .2p$ using cross-validation. Since, rank cannot be fixed in SVT, we try various values for the parameter τ to obtain the desired rank solution. Note that SVP-Newton incurs a RMSE of 0.89 for $k = 3$. In contrast, SVT achieves a RMSE of 0.98 for the same rank. We remark that SVT was able to achieve RMSE up to 0.89 but required rank 17 solution and was significantly slower in convergence because many intermediate iterates had large rank (up to around 150). We attribute the relatively poor performance of SVP and SVT as compared with ALS and SVP-Newton to the fact that the ratings matrix is not sampled uniformly, thus violating the crucial assumption of uniformly distributed samples.

References

[1] Movie lens dataset. Public dataset. URL `http://www.grouplens.org/taxonomy/term/14`.

[2] K. Arrow, L. Hurwicz, and H. Uzawa. *Studies in Linear and Nonlinear Programming*. Stanford University Press, Stanford, 1958.

[3] Robert Bell and Yehuda Koren. Scalable collaborative filtering with jointly derived neighborhood interpolation weights. In *ICDM*, pages 43–52, 2007. doi: 10.1109/ICDM.2007.90.

[4] Matthew Brand. Fast online SVD revisions for lightweight recommender systems. In *SIAM International Conference on Data Mining*, 2003.

[5] Jian-Feng Cai, Emmanuel J. Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[6] Emmanuel J. Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, December 2009.

[7] Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inform. Theory*, 56(5):2053–2080, 2009.

[8] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. In *American Control Conference, Arlington, Virginia*, 2001.

[9] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices. In *American Control Conference*, 2003.

[10] M. Fazel, E. Candes, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *Signals, Systems and Computers, 2008 42nd Asilomar Conference on*, pages 1043–1047, Oct. 2008. doi: 10.1109/ACSSC.2008.5074571.

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

[12] Donald Goldfarb and Shiqian Ma. Convergence of fixed point continuation algorithms for matrix rank minimization, 2009. Submitted.

[13] Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *ICML*, 2009.

[14] Raghunandan H. Keshavan, Sewoong Oh, and Andrea Montanari. Matrix completion from a few entries. In *ISIT'09: Proceedings of the 2009 IEEE international conference on Symposium on Information Theory*, pages 324–328, Piscataway, NJ, USA, 2009. IEEE Press. ISBN 978-1-4244-4312-3.

[15] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *KDD*, pages 426–434, 2008. doi: 10.1145/1401890.1401944.

[16] R.M. Larsen. Propack: a software for large and sparse SVD calculations. Available online. URL `http://sun.stanford.edu/rmunk/PROPACK/`.

[17] Kiryung Lee and Yoram Bresler. Admira: Atomic decomposition for minimum rank approximation, 2009.

[18] Kiryung Lee and Yoram Bresler. Guaranteed minimum rank approximation from linear observations by nuclear norm minimization with an ellipsoidal constraint, 2009.

[19] Richard B. Lehoucq, Danny C. Sorensen, and Chao Yang. *ARPACK Users' Guide: Solution of Large-Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods*. SIAM, Philadelphia, 1998.

[20] S. Ma, D. Goldfarb, and L. Chen. Fixed point and bregman iterative methods for matrix rank minimization. *To appear, Mathematical Programming Series A*, 2010.

[21] Arian Maleki. Coherence analysis of iterative thresholding algorithms. *CoRR*, abs/0904.1193, 2009.

[22] Raghu Meka, Prateek Jain, Constantine Caramanis, and Inderjit S. Dhillon. Rank minimization via online learning. In *ICML*, pages 656–663, 2008. doi: 10.1145/1390156.1390239.

[23] Raghu Meka, Prateek Jain, and Inderjit S. Dhillon. Matrix completion from power-law distributed samples. In *NIPS*, 2009.

[24] Benjamin Recht, Maryam Fazel, and Pablo A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization, 2007. To appear in SIAM Review.

[25] K.C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. Preprint, 2009. URL `http://www.math.nus.edu.sg/~matys/apg.pdf`.