Mach Learn (2017) 106:419–457
DOI 10.1007/s10994-017-5628-6

Online optimization for max-norm regularization

Jie Shen[1] · **Huan Xu**[2] · **Ping Li**[1,3]

Received: 8 November 2015 / Accepted: 19 January 2017 / Published online: 7 February 2017

Abstract The max-norm regularizer has been extensively studied in the last decade as it promotes an effective low-rank estimation for the underlying data. However, such max-norm regularized problems are typically formulated and solved in a batch manner, which prevents it from processing big data due to possible memory bottleneck. In this paper, hence, we propose an online algorithm that is scalable to large problems. In particular, we consider the matrix decomposition problem as an example, although a simple variant of the algorithm and analysis can be adapted to other important problems such as matrix completion. The crucial technique in our implementation is to reformulate the max-norm to an equivalent matrix factorization form, where the factors consist of a (possibly overcomplete) basis component and a coefficients one. In this way, we may maintain the basis component in the memory and optimize over it and the coefficients for each sample alternatively. Since the size of the basis component is independent of the sample size, our algorithm is appealing when manipulating a large collection of samples. We prove that the sequence of the solutions (i.e., the basis component) produced by our algorithm converges to a stationary point of the expected loss function asymptotically. Numerical study demonstrates encouraging results for the robustness of our algorithm compared to the widely used nuclear norm solvers.

Editor: Ulf Brefeld.

✉ Jie Shen
js2007@rutgers.edu

Huan Xu
huan.xu@isye.gatech.edu

Ping Li
pingli@stat.rutgers.edu

[1] Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA

[2] H. Milton Stewart School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA 30332, USA

[3] Department of Statistics and Biostatistics, Rutgers University, Piscataway, NJ 08854, USA

Keywords Low-rank matrix · Max-norm · Stochastic optimization · Matrix factorization

1 Introduction

In the last decade, estimating low-rank matrices has attracted increasing attention in the machine learning community owing to its successful applications in a wide range of fields including subspace clustering (Liu et al. 2010), collaborative filtering (Foygel et al. 2012) and robust dimensionality reduction (Candès et al. 2011), to name a few. Suppose that we are given an observed data matrix Z in $\mathbb{R}^{p \times n}$, i.e., n observations in p ambient dimensions, we aim to learn a prediction matrix X with a low-rank structure so as to approximate the observation. This problem, together with its many variants, typically involves minimizing a weighted combination of the residual error and a penalty for the matrix rank.

Generally speaking, it is intractable to optimize a matrix rank (Recht et al. 2010). To tackle this challenge, researchers suggested alternative convex relaxations to the matrix rank. The two most widely used convex surrogates are the nuclear norm[1] (Recht et al. 2010) and the max-norm (a.k.a. γ_2-norm) (Srebro et al. 2004). The nuclear norm is defined as the sum of the matrix singular values. Like the ℓ_1 norm in the vector case that induces sparsity, the nuclear norm was proposed as a rank minimization heuristic and was able to be formulated as a semi-definite programming (SDP) problem (Fazel et al. 2001). By combining the SDP formulation and the matrix factorization technique, Srebro et al. (2004) showed that the collaborative filtering problem can be effectively solved by optimizing a soft-margin based program. Another interesting work on the nuclear norm comes from the data compression community. In real-world applications, due to possible sensor failure and background clutter, the underlying data can easily be corrupted. In this case, estimates produced by Principal Component Analysis (PCA) may be deviated far from the true subspace (Jolliffe 2005). To handle the (gross) corruption, in the seminal work, Candès et al. (2011) proposed a new formulation termed Robust PCA (RPCA), and proved that under mild conditions, solving a convex optimization problem consisting of a nuclear norm regularization and a weighted ℓ_1 norm penalty can exactly recover the low-rank component of the underlying data even if a constant fraction of the entries are arbitrarily corrupted.

The max-norm variant was developed as another convex relaxation to the rank function (Srebro et al. 2004), where Srebro et al. formulated the max-norm regularized problem as an SDP and empirically showed the superiority to the nuclear norm. The main theoretical study on the max-norm comes from Srebro and Shraibman (2005), where Srebro and Shraibman considered collaborative filtering as an example and proved that the max-norm scheme enjoys a lower generalization error than the nuclear norm. Following these theoretical foundations, Jalali and Srebro (2012) improved the error bound for the clustering problem. Another important contribution from Jalali and Srebro (2012) is that they partially characterized the subgradient of the max-norm, which is a hard mathematical entity and cannot be fully understood to date. However, since SDP solver is not scalable, there is a large gap between the theoretical progress and the practical applicability of the max-norm. To bridge the gap, a number of follow-up works attempted to design efficient algorithms to solve max-norm regularized or constrained problems. For example, Rennie and Srebro (2005) devised a gradient-based optimization method and empirically showed promising results on large collaborative filtering datasets. Lee et al. (2010) presented large-scale optimization methods

[1] Also known as the trace norm, the Ky-Fan n-norm and the Schatten 1-norm.

for max-norm constrained and max-norm regularized problems and showed a convergence to stationary point.

Nevertheless, algorithms presented in prior works (Srebro et al. 2004; Rennie and Srebro 2005; Lee et al. 2010; Orabona et al. 2012) require to access all the data when the objective function involves a max-norm regularization. In the large-scale setting, the applicability of such batch optimization methods will be hindered by the memory bottleneck. In this paper, henceforth, we propose an online algorithm to solve max-norm regularized problems. The main advantage of online algorithms is that the memory cost is independent of the sample size, which makes it a good fit for the *big data* era.

To be more detailed, we are interested in a general max-norm regularized matrix decomposition (MRMD) problem. Suppose that the observed data matrix Z can be decomposed into a low-rank component X and some structured noise E, we aim to simultaneously and accurately estimate the two components, by solving the following convex program:

$$\text{(MRMD)} \quad \min_{X,E} \quad \frac{1}{2}\left\|Z - X - E\right\|_F^2 + \frac{\lambda_1}{2}\left\|X\right\|_{\max}^2 + \lambda_2 h(E). \tag{1.1}$$

Here, $\|\cdot\|_F$ denotes the Frobenius norm which is a commonly used metric for evaluating the residual, $\|\cdot\|_{\max}$ is the max-norm (which promotes low-rankness), and λ_1 and λ_2 are two non-negative parameters. $h(E)$ is some (convex) regularizer that can be adapted to various kinds of noise. We require that it can be represented as a summation of column norms. Formally, there exists some regularizer $\tilde{h}(\cdot)$, such that

$$h(E) = \sum_{i=1}^{n} \tilde{h}(\boldsymbol{e}_i), \tag{1.2}$$

where $\boldsymbol{e}_i$ is the ith column of E. Classical examples include:

- $\|E\|_1$. That is, the ℓ_1 norm of the matrix E seen as a long vector, which is used to handle sparse corruption. In this case, $\tilde{h}(\cdot)$ is the ℓ_1 vector norm. Note that when equipped with this norm, the above problem reduces to the well-known RPCA formulation (Candès et al. 2011), but with the nuclear norm being replaced by the max-norm.
- $\|E\|_{2,1}$. This is defined as the summation of the ℓ_2 column norms, which is effective when a small fraction of the samples are contaminated (recall that each column of Z is a sample). The matrix $\ell_{2,1}$ norm is typically used to handle outliers and interestingly, the above program becomes Outlier PCA (Xu et al. 2013) in this case.
- $\|E\|_F^2$ or $E = 0$. The formulation of (1.1) works as a large-margin based program, with the hinge loss replaced by the squared loss (Srebro et al. 2004).

Hence, (MRMD) (1.1) is general enough and our algorithmic and theoretical results hold for such a general form, covering important problems including max-norm regularized RPCA, max-norm regularized Outlier PCA and maximum margin matrix factorization. Furthermore, with a careful design, the above formulation (1.1) can be extended to address the matrix completion problem (Candès and Recht 2009), as we will show in Sect. 5.

Considering the connection between max-norm and nuclear norm, one might be interested in an alternative formulation as follows:

$$\min_{X,E} \quad \frac{1}{2}\left\|Z - X - E\right\|_F^2 + \frac{\lambda_1'}{2}\left\|X\right\|_{\max} + \lambda_2 h(E). \tag{1.3}$$

First, we would like to point out that the above formulation is equivalent to (1.1), in the sense that if we choose proper parameter λ_1' for (1.3) and some parameter λ_1 for (1.1), they produce same solutions. To see this, we note that (1.3) is equivalent to the following constrained program:

$$\min_{X,E} \frac{1}{2} \|Z - X - E\|_F^2 + \lambda_2 h(E), \quad \text{s.t. } \|X\|_{\max} \leq \kappa,$$

for some parameter κ. Taking the square on both sides of the inequality constraint gives

$$\min_{X,E} \frac{1}{2} \|Z - X - E\|_F^2 + \lambda_2 h(E), \quad \text{s.t. } \|X\|_{\max}^2 \leq \kappa^2.$$

Again, we know that for some proper choice of λ_1, the above program is equivalent to (1.1). The reason we choose (1.1) is for a convenient computation of the solution. We defer a more detailed discussion to Sect. 3.

1.1 Contributions

In summary, our main contributions is two-folds: (1) We are the first to develop an online algorithm to solve a family of max-norm regularized problems (1.1), which admits a wide range of applications in machine learning. We also show that our approach can be used to solve other popular max-norm regularized problems such as matrix completion. (2) We prove that the sequence of solutions produced by our algorithm converges to a stationary point of the expected loss function asymptotically (see Sect. 4).

Compared to our earlier work (Shen et al. 2014), the formulation (1.1) considered here is more general and a complete proof is provided. In addition, we illustrate by an extensive study on the subspace recovery task that the max-norm always performs better than the nuclear norm in terms of convergence rate and robustness.

1.2 Related works

Here we discuss some relevant works in the literature. Most previous works on max-norm focused on showing that it is empirically superior to the nuclear norm in real-world problems, such as collaborative filtering (Srebro et al. 2004), clustering (Jalali and Srebro 2012) and hamming embedding (Neyshabur et al. 2014). Other works, for instance, Salakhutdinov and Srebro (2010) studied the influence of data distribution with the max-norm regularization and observed good performance even when the data are sampled non-uniformly. There are also interesting works which investigated the connection between the max-norm and the nuclear norm. A comprehensive study on this problem, in the context of collaborative filtering, can be found in Srebro and Shraibman (2005), which established and compared the generalization bound for the nuclear norm regularization and the max-norm, showing that the latter one results in a tighter bound. More recently, Foygel et al. (2012) attempted to unify them to gain insightful perspective.

Also in line with this work is matrix decomposition. As we mentioned, when we penalize the noise E with ℓ_1 matrix norm, it reverts to the well known RPCA formulation (Candès et al. 2011). The only difference is that Candès et al. (2011) analyzed the RPCA problem with the nuclear norm, while (1.1) employs the max-norm. Owing to the explicit form of the subgradient of the nuclear norm, Candès et al. (2011) established a dual certificate for the success of their formulation, which facilitates their theoretical analysis. In contrast, the max-norm is a much harder mathematical entity (even its subgradient has not been fully characterized). Henceforth, it still remains challenging to understand the behavior of the max-norm regularizer in the general setting (1.1). Studying the conditions for the exact recovery of MRMD is out of the scope of this paper. We leave this as a future work.

From a high level, the goal of this paper is similar to that of Feng et al. (2013). Motivated by the celebrated RPCA problem (Candès et al. 2011; Xu et al. 2013, 2012), Feng et al. (2013) developed an online implementation for the nuclear-norm regularized matrix decomposition. Yet, since the max-norm is a more complicated mathematical entity, new techniques and insights are needed in order to develop online methods for the max-norm regularization. For example, after converting the max-norm to its matrix factorization form, the data are still coupled and we propose to transform the problem to a constrained one for stochastic optimization.

The main technical contribution of this paper is converting max-norm regularization to an appropriate matrix factorization problem that is amenable to online implementation. Compared to Mairal et al. (2010) which also studies online matrix factorization, our formulation contains an additional structured noise that brings the benefit of robustness to contamination. Some of our proof techniques are also different. For example, to prove the convergence of the dictionary and to well define their problem, Mairal et al. (2010) assumed that the magnitude of the learned dictionary is constrained. In contrast, we *prove* that the optimal basis is uniformly bounded, and hence our problem is naturally well-defined.

Our algorithm can be viewed as a majorization-minimization scheme, for which Mairal (2013) derived a general analysis on the convergence behavior. However, we find that Algorithm 1 in Mairal (2013) requires the knowledge of the Lipschitz constant to obtain a surrogate function. In our work, we use a suboptimal solution to derive the surrogate function (see Step 3 and Step 5 in our Algorithm 1 to be introduced). Due to such a different mechanism, it remains an open question whether one can apply their algorithm and theoretical analysis to the problem considered here. It is also worth mentioning that in order to establish their theoretical results, Mairal (2013) *assumed* that the iterates and the empirical loss function are uniformly bounded (see Assumption (C) and Assumption (D) therein). For our problem, we can virtually prove this property (see Proposition 3 and Corollary 1 to follow). Finally, we note that our algorithm is different from block coordinate descent, see, e.g., Wang and Banerjee (2014). In fact, block coordinate descent randomly and independently picks a mini-batch of samples and updates a block variable, whereas we in each iteration update only the variables associated with the revealed sample. Another key difference is that Wang and Banerjee (2014) considered a strongly convex objective function, while we are working with a non-convex case.

1.3 Roadmap

The rest of the paper is organized as follows. Section 2 begins with the problem setting, followed by a reformulation of the MRMD problem that is amenable for online optimization. Section 3 then elaborates the online implementation of MRMD and Sect. 4 establishes the convergence guarantee under some mild assumptions. In Sect. 5, we show that our framework can easily be extended to other max-norm regularized problems, such as matrix completion. Numerical performance of the proposed algorithm is presented in Sect. 6. Finally, we conclude this paper in Sect. 7. All the proofs are deferred to the “Appendix”.

1.4 Notation

Before delivering the algorithm and the analysis, let us first instate several pieces of notation that are involved throughout the paper. We use bold lowercase letters, e.g., $\boldsymbol{v}$, to denote a column vector. The ℓ_1 norm and ℓ_2 norm of a vector $\boldsymbol{v}$ are denoted by $\|\boldsymbol{v}\|_1$ and $\|\boldsymbol{v}\|_2$, respectively. Capital letters, such as M, are used to denote matrices. In particular, the letter I_n is reserved for the n-by-n identity matrix. For a matrix M, the ith row and jth column are written as $\boldsymbol{m}(i)$ and $\boldsymbol{m}_j$, respectively, and the (i, j)th entry is denoted by m_{ij}. There are four matrix norms that will be heavily used in the paper: $\|M\|_F$ for the Frobenius norm, $\|M\|_1$ for the ℓ_1 matrix norm seen as a long vector, $\|M\|_{\max}$ for the max-norm induced by the product of $\ell_{2,\infty}$ norm on the factors of M. Here, the $\ell_{2,\infty}$ norm is defined as the maximum ℓ_2 row norm. The trace of a square matrix M is denoted as $\mathrm{Tr}(M)$. Finally, for a positive integer n, we use $[n]$ to denote the integer set $\{1, 2, \ldots, n\}$.

2 Problem setup

We are interested in developing an online algorithm for the MRMD problem (1.1) so as to mitigate the memory issue. To this end, we utilize the following definition of the max-norm (Srebro et al. 2004):

$$\|X\|_{\max} \overset{\text{def}}{=} \min_{L,R} \left\{ \|L\|_{2,\infty} \cdot \|R\|_{2,\infty} :\ X = LR^\top, L \in \mathbb{R}^{p\times d}, R \in \mathbb{R}^{n\times d} \right\}, \tag{2.1}$$

where d is an upper bound on the intrinsic dimension of the underlying data. Plugging the above back to (1.1), we obtain an equivalent form:

$$\min_{L,R,E} \ \frac{1}{2} \left\| Z - LR^\top - E \right\|_F^2 + \frac{\lambda_1}{2} \|L\|_{2,\infty}^2 \|R\|_{2,\infty}^2 + \lambda_2 h(E). \tag{2.2}$$

In this paper, if not specified, "equivalent" means we do not change the optimal value of the objective function. Intuitively, the variable L serves as a (possibly overcomplete) basis for the clean data while correspondingly, the variable R works as a coefficients matrix with each row being the coefficients for each sample (recall that we organize the observed samples in a column-wise manner). In order to make the new formulation (2.2) equivalent to MRMD (1.1), the quantity of d should be sufficiently large due to (2.1).

At a first sight, the problem can only be optimized in a batch manner for which the memory cost is prohibitive. To see this, note that we are considering the regime of $d < p \ll n$ and the size of the coefficients R is proportional to n. In order to optimize the above program over the variable R, we have to compute the gradient with respect to it. Recall that the $\ell_{2,\infty}$ norm counts the largest ℓ_2 row norm of R, hence coupling all the samples (each row of R associates with a sample).

Fortunately, we have the following proposition that alleviates the inter-dependency among the rows of R, hence facilitating an online algorithm where the rows of R can be optimized sequentially.

Proposition 1 *Problem* (2.2) *is equivalent to the following constrained program:*

$$\min_{L,R,E} \frac{1}{2} \left\| Z - LR^\top - E \right\|_F + \frac{\lambda_1}{2} \|L\|_{2,\infty}^2 + \lambda_2 h(E), \quad \text{s.t. } \|R\|_{2,\infty}^2 \le 1. \tag{2.3}$$

Moreover, there exists an optimal solution (L^*, R^*, E^*) attained at the boundary of the feasible set, i.e., $\|R^*\|_{2,\infty}^2$ is equal to the unit.

Remark 1 Proposition 1 is crucial for the online implementation. It states that our primal MRMD problem (1.1) can be transformed to an equivalent constrained program (2.3) where the coefficients of *each individual* sample (i.e., a row of the matrix R) is *uniformly and separately* constrained.

Consequently, we can, equipped with Proposition 1, rewrite the original problem in an online fashion, with each sample being separately processed:

$$\min_{L,R,E} \frac{1}{2}\sum_{i=1}^{n} \|\boldsymbol{z}_i - L\boldsymbol{r}_i - \boldsymbol{e}_i\|_2^2 + \frac{\lambda_1}{2}\|L\|_{2,\infty}^2 + \lambda_2 \sum_{i=1}^{n} \tilde{h}(\boldsymbol{e}_i), \text{ s.t. } \|\boldsymbol{r}_i\|_2^2 \le 1, \ \forall\, i \in [n], \tag{2.4}$$

where $\boldsymbol{z}_i$ is the ith observation, $\boldsymbol{r}_i$ is the coefficients and $\boldsymbol{e}_i$ is some structured error penalized by the (convex) regularizer $\tilde{h}(\cdot)$ (recall that we require $h(E)$ can be decomposed column-wisely). Merging the first and third term above gives a compact form:

$$\min_{L} \min_{R,E} \sum_{i=1}^{n} \tilde{\ell}(\boldsymbol{z}_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) + \frac{\lambda_1}{2}\|L\|_{2,\infty}^2, \quad \text{s.t. } \|\boldsymbol{r}_i\|_2^2 \le 1, \ \forall i \in [n], \tag{2.5}$$

where

$$\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e}) \overset{\text{def}}{=} \frac{1}{2}\|\boldsymbol{z} - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}). \tag{2.6}$$

This is indeed equivalent to optimizing (i.e., minimizing) the empirical loss function:

$$\min_{L} f_n(L), \tag{2.7}$$

where

$$f_n(L) \overset{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \ell(\boldsymbol{z}_i, L) + \frac{\lambda_1}{2n}\|L\|_{2,\infty}^2, \tag{2.8}$$

and

$$\ell(\boldsymbol{z}, L) = \min_{\boldsymbol{r},\boldsymbol{e},\|\boldsymbol{r}\|_2^2 \le 1} \tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e}). \tag{2.9}$$

Note that by Proposition 1, as long as the quantity of d is sufficiently large, the program (2.7) is equivalent to the primal formulation (1.1), in the sense that both of them could attain the same minimum. Compared to MRMD (1.1), which is solved in a batch manner by prior works, the formulation (2.7) paves a way for stochastic optimization procedure since all the samples are decoupled.

3 Algorithm

Based on the derivation in the preceding section, we are now ready to present our online algorithm to solve the MRMD problem (1.1). The implementation is outlined in Algorithm 1. Here we briefly explain the underlying intuition. We optimize the coefficients $\boldsymbol{r}$, the structured noise $\boldsymbol{e}$ and the basis L in an alternating manner, with only the basis L and two accumulation matrices being kept in memory. At the tth iteration, given the basis L_{t-1} produced by the previous iteration, we can optimize (2.9) by examining the Karush–Kuhn–Tucker (KKT) conditions. To obtain a new iterate L_t, we then minimize the following objective function:

Algorithm 1 Online Max-Norm Regularized Matrix Decomposition

```
Require: $Z \in \mathbb{R}^{p \times n}$ (observed samples), parameters $\lambda_1$ and $\lambda_2$, $L_0 \in \mathbb{R}^{p \times d}$ (initial basis), zero matrices $A_0 \in \mathbb{R}^{d \times d}$ and $B_0 \in \mathbb{R}^{p \times d}$.
Ensure: Optimal basis $L_n$.
for $t = 1$ to $n$ do
    Access the $t$th sample $z_t$.
    Compute the coefficient and noise:
        $\{\boldsymbol{r}_t, \boldsymbol{e}_t\} = \operatorname*{arg\,min}_{\boldsymbol{r}, \boldsymbol{e}, \|\boldsymbol{r}\|_2^2 \le 1} \tilde{\ell}(z_t, L_{t-1}, \boldsymbol{r}, \boldsymbol{e}).$
    Compute the accumulation matrices $A_t$ and $B_t$:
        $A_t \longleftarrow A_{t-1} + \boldsymbol{r}_t \boldsymbol{r}_t^\top,$
        $B_t \longleftarrow B_{t-1} + (z_t - \boldsymbol{e}_t)\, \boldsymbol{r}_t^\top.$
    Compute the basis $L_t$ by optimizing the surrogate function (3.1):
        $L_t = \operatorname*{arg\,min}_{L} \frac{1}{t} \sum_{i=1}^{t} \tilde{\ell}(z_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2$
        $= \operatorname*{arg\,min}_{L} \frac{1}{t} \left( \frac{1}{2} \operatorname{Tr}\left(L^\top L A_t\right) - \operatorname{Tr}\left(L^\top B_t\right) \right) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2 .$
end for
```

$$g_t(L) \overset{\text{def}}{=} \frac{1}{t} \sum_{i=1}^{t} \tilde{\ell}(z_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2 , \tag{3.1}$$

where $\{\boldsymbol{r}_i\}_{i=1}^t$ and $\{\boldsymbol{e}_i\}_{i=1}^t$ are already on hand. It can be verified that (3.1) is a surrogate function of the empirical loss $f_t(L)$ (2.8), since the obtained $\boldsymbol{r}_i$'s and $\boldsymbol{e}_i$'s are suboptimal. Interestingly, instead of recording all the past $\boldsymbol{r}_i$'s and $\boldsymbol{e}_i$'s, we only need to store two accumulation matrices whose sizes are independent of n, as shown in Algorithm 1. In the sequel, we elaborate each step.

3.1 Update the coefficients and noise

Given a sample z and a basis L, we are able to estimate the optimal coefficients $\boldsymbol{r}$ and the noise $\boldsymbol{e}$ by minimizing $\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e})$. That is, we are to solve the following program:

$$\min_{\boldsymbol{r}, \boldsymbol{e}} \frac{1}{2} \|z - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}), \quad \text{s.t. } \|\boldsymbol{r}\|_2 \le 1. \tag{3.2}$$

We notice that the constraint only involves the variable $\boldsymbol{r}$, and in order to optimize $\boldsymbol{r}$, we only need to consider the residual term in the objective function. This motivates us to employ a block coordinate descent algorithm. Namely, we alternatively optimize one variable with the other fixed, until some stopping criteria is fulfilled. In our implementation, when the difference between the current and the previous iterate is smaller than 10^{-6}, or the number of iterations exceeds 100, our algorithm will terminate and return the optimum.

3.1.1 Optimize the coefficients $\boldsymbol{r}$

Now it remains to show how to compute a new iterate for one variable when the other one is fixed. According to Bertsekas (1999), when the objective function is strongly convex with respect to (w.r.t.) each block variable, we are guaranteed that the block coordinate minimization algorithm converges. In our case, we observe that such a condition holds for $\boldsymbol{e}$ but not necessary for $\boldsymbol{r}$. In fact, the strong convexity w.r.t. $\boldsymbol{r}$ holds if and only if the basis L is with full rank. When L is not full rank, we may compute the Moore Penrose pseudo inverse to solve $\boldsymbol{r}$. However, for computational efficiency, we append a small jitter $\frac{\epsilon}{2}\|\boldsymbol{r}\|_2^2$ to the objective if necessary, so as to guarantee the convergence ($\epsilon = 0.01$ in our experiments). In this way, we obtain a *potentially* admissible iterate for $\boldsymbol{r}$ as follows:

$$\boldsymbol{r}_0 = (L^\top L + \epsilon I_d)^{-1} L^\top (\boldsymbol{z} - \boldsymbol{e}). \tag{3.3}$$

Here, ϵ is set to be zero if and only if L is full rank.

Next, we examine if $\boldsymbol{r}_0$ violates the inequality constraint in (3.2). If it happens to be a feasible solution, i.e., $\|\boldsymbol{r}_0\|_2 \leq 1$, we have found the new iterate for $\boldsymbol{r}$. Otherwise, we conclude that the optimum of $\boldsymbol{r}$ must be attained on the boundary of the feasible set, i.e., $\|\boldsymbol{r}\|_2 = 1$, for which the minimizer can be computed by the method of Lagrangian multipliers:

$$\max_\eta \min_{\boldsymbol{r}} \frac{1}{2}\|\boldsymbol{z} - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \frac{\eta}{2}\left(\|\boldsymbol{r}\|_2^2 - 1\right), \quad \text{s.t. } \eta > 0, \quad \|\boldsymbol{r}\|_2 = 1. \tag{3.4}$$

By differentiating the objective function with respect to $\boldsymbol{r}$, we have

$$\boldsymbol{r} = \left(L^\top L + \eta I_d\right)^{-1} L^\top (\boldsymbol{z} - \boldsymbol{e}). \tag{3.5}$$

The following argument helps us to efficiently search the optimal solution.

Proposition 2 *Let $\boldsymbol{r}$ be given by (3.5), where L, $\boldsymbol{z}$ and $\boldsymbol{e}$ are assumed to be fixed. Then, the ℓ_2 norm of $\boldsymbol{r}$ is strictly monotonically decreasing with respect to the quantity of η.*

Proof For simplicity, let us denote

$$\boldsymbol{r}(\eta) = \left(L^\top L + \eta I_d\right)^{-1} \boldsymbol{b},$$

where $\boldsymbol{b} = L^\top(\boldsymbol{z} - \boldsymbol{e})$ is a fixed vector. Suppose we have a full singular value decomposition (SVD) on $L = USV^\top$, where the singular values $\{s_1, s_2, \ldots, s_p\}$ (i.e., the diagonal elements in S) are arranged in a decreasing order and at most d number of them are non-zero. Substituting L with its SVD, we obtain the squared ℓ_2 norm for $\boldsymbol{r}(\eta)$:

$$\|\boldsymbol{r}(\eta)\|_2^2 = \boldsymbol{b}^\top \left(VS^2V^\top + \eta I_d\right)^{-2} \boldsymbol{b} = \boldsymbol{b}^\top V S_\eta V^\top \boldsymbol{b},$$

where S_η is a diagonal matrix whose ith diagonal element equals $(s_i^2 + \eta)^{-2}$.

For any two entities $\eta_1 > \eta_2$, it is easy to see that the matrix $S_{\eta_1} - S_{\eta_2}$ is negative definite. Hence, it always holds that

$$\|\boldsymbol{r}(\eta_1)\|_2^2 - \|\boldsymbol{r}(\eta_2)\|_2^2 = \boldsymbol{b}^\top V(S_{\eta_1} - S_{\eta_2})V^\top \boldsymbol{b} < 0,$$

which concludes the proof.

The above proposition offers an efficient computation scheme, i.e., bisection method, for searching the optimal $\boldsymbol{r}$ as well as the dual variable η. To be more detailed, we can maintain a

Algorithm 2 Bisection Method for Problem (3.4)

```
Require: $L \in \mathbb{R}^{p \times d}$, $\boldsymbol{z} \in \mathbb{R}^p$, $\boldsymbol{e} \in \mathbb{R}^p$.
Ensure: Optimal primal and dual pair $(\boldsymbol{r}, \eta)$.
Initialize the lower bound $\eta_1 = 0$ and the upper bound $\eta_2$ large enough such that $\|\boldsymbol{r}(\eta_2)\|_2 \le 1$.
repeat
    Compute the middle point:
        $\eta \leftarrow \frac{1}{2}(\eta_1 + \eta_2).$
    if $\|\boldsymbol{r}(\eta)\|_2 < 1$ then
        Update $\eta_2$:
            $\eta_2 \leftarrow \eta.$
    else
        Update $\eta_1$:
            $\eta_1 \leftarrow \eta.$
    end if
until $\|\boldsymbol{r}\|_2 = 1$
```

lower bound η_1 and an upper bound η_2, such that $\|\boldsymbol{r}(\eta_1)\|_2 \ge 1$ and $\|\boldsymbol{r}(\eta_2)\|_2 \le 1$. According to the monotonic property shown in Proposition 2, the optimal η must fall into the interval $[\eta_1, \eta_2]$. By evaluating the value of $\|\boldsymbol{r}\|_2$ at the middle point $(\eta_1 + \eta_2)/2$, we can sequentially shrink the interval until $\|\boldsymbol{r}\|_2$ is close or equal to one. Note that we can initialize η_1 with zero (since $\left\|\boldsymbol{r}_0\right\|_2 > 1$ implies the optimal $\eta^* > \epsilon \ge 0$). The bisection routine is summarized in Algorithm 2.

3.1.2 *Optimize the noise* $\boldsymbol{e}$

We have clarified the technique used for solving $\boldsymbol{r}$ in Problem (3.2) when $\boldsymbol{e}$ is fixed. Now let us turn to the phase where $\boldsymbol{r}$ is fixed and we want to find the optimal $\boldsymbol{e}$. Since $\boldsymbol{e}$ is an unconstrained variable, generally speaking, it is much easier to solve, although one may employ different strategies for various regularizers $\tilde{h}(\cdot)$. Here, we discuss the solutions for popular choices of the regularizer.

1. $\tilde{h}(\boldsymbol{e}) = \|\boldsymbol{e}\|_1$. The ℓ_1 regularizer results in a closed form solution for $\boldsymbol{e}$ as follows:
$$\boldsymbol{e} = \mathcal{S}_{\lambda_2}[\boldsymbol{z} - L\boldsymbol{r}], \tag{3.6}$$
where $\mathcal{S}_{\lambda_2}[\cdot]$ is the soft-thresholding operator (Donoho 1995).
2. $\tilde{h}(\boldsymbol{e}) = \|\boldsymbol{e}\|_2$. The solution in this case can be characterized as follows (see, for example, Liu et al. 2010):
$$\boldsymbol{e} = \begin{cases} \frac{\|\boldsymbol{z} - L\boldsymbol{r}\|_2}{\|\boldsymbol{z} - L\boldsymbol{r}\|_2 - \lambda_2}(\boldsymbol{z} - L\boldsymbol{r}), & \text{if } \lambda_2 < \|\boldsymbol{z} - L\boldsymbol{r}\|_2\,, \\ \boldsymbol{0}, & \text{otherwise.} \end{cases} \tag{3.7}$$

Finally, for completeness, we summarize the routine for updating the coefficients and the noise in Algorithm 3. The readers may refer to the preceding paragraphs for details.

Algorithm 3 The Coefficients and Noise Update (Problem (3.2))

```
Require: $L \in \mathbb{R}^{p \times d}$, $z \in \mathbb{R}^p$, parameter $\lambda_2$ and a small jitter $\epsilon$.
Ensure: Optimal $\boldsymbol{r}$ and $\boldsymbol{e}$.
Initialize $\boldsymbol{e} = \boldsymbol{0}$.
repeat
    Compute the potential solution $\boldsymbol{r}_0$ given in (3.3).
    if $\|\boldsymbol{r}_0\|_2 \le 1$ then
        Update $\boldsymbol{r}$ with
            $$\boldsymbol{r} = \boldsymbol{r}_0,$$
    else
        Update $\boldsymbol{r}$ by Algorithm 2.
    end if
    Update the noise $\boldsymbol{e}$.
until convergence
```

3.2 Update the basis

With all the past filtration $\mathcal{F}_t = \{z_i, \boldsymbol{r}_i, \boldsymbol{e}_i\}_{i=1}^t$ on hand, we are able to compute a new basis L_t by minimizing the surrogate function (3.1). That is, we are to solve the following program:

$$\min_L \quad \frac{1}{t}\sum_{i=1}^t \tilde{\ell}(z_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) + \frac{\lambda_1}{2t}\|L\|_{2,\infty}^2. \tag{3.8}$$

By a simple expansion, for any $i \in [t]$, we have

$$\tilde{\ell}(z_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) = \frac{1}{2}\operatorname{Tr}\left(L^\top L \boldsymbol{r}_i \boldsymbol{r}_i^\top\right) - \operatorname{Tr}\left(L^\top (z_i - \boldsymbol{e}_i)\boldsymbol{r}_i^\top\right) + \frac{1}{2}\|z_i - \boldsymbol{e}_i\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}_i). \tag{3.9}$$

Substituting back into (3.8), putting $A_t = \sum_{i=1}^t \boldsymbol{r}_i\boldsymbol{r}_i^\top$, $B_t = \sum_{i=1}^t (z_i - \boldsymbol{e}_i)\boldsymbol{r}_i^\top$ and removing constant terms, we obtain

$$L_t = \arg\min_L \frac{1}{t}\left(\frac{1}{2}\operatorname{Tr}\left(L^\top L A_t\right) - \operatorname{Tr}\left(L^\top B_t\right)\right) + \frac{\lambda_1}{2t}\|L\|_{2,\infty}^2. \tag{3.10}$$

In order to derive the optimal solution, firstly, we need to characterize the subgradient of the squared $\ell_{2,\infty}$ norm. In fact, let Q be a positive semi-definite diagonal matrix, such that $\operatorname{Tr}(Q) = 1$. Denote the set of row index which attains the maximum ℓ_2 row norm of L by $\mathcal{I}$. In this way, the subgradient of $\frac{1}{2}\|L\|_{2,\infty}^2$ is given by

$$\partial\left(\frac{1}{2}\|L\|_{2,\infty}^2\right) = QL,\ \ Q_{ii} \neq 0 \text{ if and only if } i \in \mathcal{I},\ \ Q_{ij} = 0 \text{ for } i \neq j. \tag{3.11}$$

Equipped with the subgradient, we may apply block coordinate descent to update each column of L sequentially. We assume that the objective function (3.10) is strongly convex w.r.t. L, implying that the block coordinate descent scheme can always converge to the global optimum (Bertsekas 1999).

We summarize the update procedure in Algorithm 4. In practice, we find that after revealing a large number of samples, performing one-pass update for each column of L is sufficient to guarantee a desirable accuracy, which matches the observation in Mairal et al. (2010).

Algorithm 4 The Basis Update

```
Require: $L \in \mathbb{R}^{p \times d}$ in the previous iteration, accumulation matrix $A$ and $B$, parameter $\lambda_1$.
Ensure: Optimal basis $L$ (updated).
repeat
    Compute the subgradient of $\frac{1}{2}\|L\|_{2,\infty}^2$:
        $U = \partial\left(\frac{1}{2}\|L\|_{2,\infty}^2\right)$.
    for $j = 1$ to $d$ do
        Update the $j$th column:
            $\boldsymbol{l}_j \leftarrow \boldsymbol{l}_j - \frac{1}{a_{jj}}\left(L\boldsymbol{a}_j - \boldsymbol{b}_j + \lambda_1 \boldsymbol{u}_j\right)$.
    end for
until convergence
```

As we discussed in Sect. 1, one may prefer the formulation (1.3)–(1.1), although in some sense they are equivalent. It is worth mentioning that our algorithm can easily be tailored to solve (1.3) by modifying Step 5 of Algorithm 1 as follows:

$$L_t = \arg\min_{L} \frac{1}{t}\left(\frac{1}{2}\operatorname{Tr}\left(L^\top L A_t\right) - \operatorname{Tr}\left(L^\top B_t\right)\right) + \frac{\lambda_1}{2t}\|L\|_{2,\infty}. \tag{3.12}$$

Again, we are required to derive the optimal solution by examining the subgradient of the last term, which is given by

$$\partial\|L\|_{2,\infty} = QW,\; q_{ii} \neq 0 \text{ if and only if } i \in \mathcal{I},\; q_{ij} = 0 \text{ for } i \neq j, \tag{3.13}$$

where each row of W is as follows:

$$\boldsymbol{w}(i) = \frac{1}{\|\boldsymbol{l}(i)\|_2}\boldsymbol{l}(i),\; \forall\, 1 \le i \le p. \tag{3.14}$$

3.3 Memory and computational cost

As one of the main contributions of this paper, our OMRMD algorithm (i.e., Algorithm 1) is appealing for large-scale problems (the regime $d < p \ll n$) since the memory cost is independent of n. To see this, note that when computing the optimal coefficients and noise, only $\boldsymbol{z}_t$ and L_{t-1} are accessed, which costs $O(pd)$ memory. To store the accumulation matrix A_t, we need $O(d^2)$ memory while that for B_t is $O(pd)$. Finally, we find that only A_t and B_t are needed for the computation of the new iterate L_t. Therefore, the total memory cost of OMRMD is $O(pd)$, i.e., independent of n. In contrast, the SDP formulation introduced by Srebro et al. (2004) requires $O((p+n)^2)$ memory usage, the local-search heuristic algorithm (Rennie and Srebro 2005) needs $O(d(p+n))$ and no convergence guarantee was derived. Even for a recently proposed algorithm (Lee et al. 2010), they require to store the entire data matrix and thus the memory cost is $O(pn)$.

In terms of computational efficiency, our algorithm can be fast. One may have noticed that the computation is dominated by solving Problem (3.2). The computational complexity of (3.5) involves an inverse of a $d \times d$ matrix followed by a matrix-matrix and a matrix-vector multiplication, totally $O(pd^2)$. For the basis update, obtaining a subgradient of the squared $\ell_{2,\infty}$ norm is $O(pd)$ since we need to calculate the ℓ_2 norm for all rows of L followed by a multiplication with a diagonal matrix (see (3.11)). A one-pass update for the columns of L, as shown in Algorithm 4 costs $O(pd^2)$. Note that the quadratic dependency on d is acceptable in the low-rank setting.

4 Theoretical analysis and proof sketch

In this section we present our main theoretical result regarding the validity of the proposed algorithm. We first discuss some necessary assumptions.

4.1 Assumptions

- $(A1)$ The observed samples are independent and identically distributed (i.i.d.) with a compact support $\mathcal{Z}$. This is a very common scenario in real-world applications.
- $(A2)$ The surrogate functions $g_t(L)$ in (3.1) are strongly convex. In particular, we assume that the smallest singular value of the positive semi-definite matrix $\frac{1}{t}A_t$ defined in Algorithm 1 is not smaller than some positive constant β_1.
- $(A3)$ The minimizer for (2.9) is unique. Notice that $\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e})$ is strongly convex w.r.t. $\boldsymbol{e}$ and convex w.r.t. $\boldsymbol{r}$. We can enforce this assumption by adding a jitter $\frac{\epsilon}{2}\|\boldsymbol{r}\|_2^2$ to the objective function, where ϵ is a small positive constant.

4.2 Main results

It is easy to see that Algorithm 1 is devised to optimize the empirical loss function (2.8). In stochastic optimization, we are mainly interested in the expected loss function, which is defined as the averaged loss incurred when the number of samples goes to infinity. If we assume that each sample is independently and identically distributed (i.i.d.), we have

$$f(L) \overset{\text{def}}{=} \lim_{n\to\infty} f_n(L) = \mathbb{E}_{\boldsymbol{z}}[\ell(\boldsymbol{z}, L)]. \tag{4.1}$$

The main theoretical result of this work is stated as follows.

Theorem 1 (Convergence to a stationary point of the expected loss function) *Let $\{L_t\}_{t=1}^{\infty}$ be the sequence of solutions produced by Algorithm 1. Then, the sequence converges to a stationary point of the expected loss function (4.1) when t tends to infinity.*

Remark 2 The theorem establishes the validity of our algorithm. Note that on one hand, the transformation (2.1) facilitates an amenable way for the online implementation of the max-norm. On the other hand, due to the non-convexity of our new formulation (2.3), it is generally hard to desire a local, or a global minimizer (Bertsekas 1999). Although Burer and Monteiro (2005) showed that any local minimum of an SDP is also the global optimum under some conditions (note that the max-norm problem can be transformed to an SDP (Srebro et al. 2004), it is not clear how to determine that a solution is a local optimum or a stationary point. Very recently, Bhojanapalli et al. (2016) showed that global convergence is possible for a family of batch methods. Yet, it is not clear how to apply their results in the stochastic setting. From the empirical study in Sect. 6, we find that the solutions produced by our algorithm always converge to the global optimum when the samples are drawn from a i.i.d. Gaussian distribution.

4.3 Proof outline

The essential tools for our analysis are from stochastic approximation (Bottou 1998) and asymptotic statistics (Vaart 2000). There are four key stages in our proof and one may find the full proof in "Appendix".

Stage I We first show that all the stochastic variables $\{L_t, \boldsymbol{r}_t, \boldsymbol{e}_t\}_{t=1}^{\infty}$ are uniformly bounded. The property is crucial because it justifies that the problem we are solving is well-defined. Also, the uniform boundedness will be heavily used for deriving subsequent important results, e.g., the Lipschitz property of the surrogate function.

Proposition 3 (Uniform bound of all stochastic variables) *Let $\{\boldsymbol{r}_t, \boldsymbol{e}_t, L_t\}_{t=1}^{\infty}$ be the sequence of the solutions produced by Algorithm 1. Then,*

1. *For any $t > 0$, the optimal solutions $\boldsymbol{r}_t$ and $\boldsymbol{e}_t$ are uniformly bounded.*
2. *For any $t > 0$, the accumulation matrices $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$ are uniformly bounded.*
3. *There exists a compact set $\mathcal{L}$, such that for any $t > 0$, we have $L_t \in \mathcal{L}$.*

Proof (Sketch) The uniform bound of $\boldsymbol{e}_t$ follows by constructing a trivial solution $(\mathbf{0}, \mathbf{0})$ for (2.6), which results in an upper bound for the optimum of the objective function. Notably, the upper bound here only involves a quantity on $\|z_t\|_2$, which is assumed to be uniformly bounded. Since $\boldsymbol{r}_t$ is always upper bounded by the unit, the first claim follows. The second claim follows immediately by combining the first claim and Assumption $(A1)$. In order to show that L_t is uniformly bounded, we utilize the first order optimality condition of the surrogate (3.1). Since $\frac{1}{t}A_t$ is positive definite, we can represent L_t in terms of $\frac{1}{t}B_t$, U_t and the inverse of $\frac{1}{t}A_t$, where U_t is the subgradient, whose Frobenius norm is in turn bounded by that of L_t. Hence, it follows that L_t can be uniformly bounded.

Remark 3 Note that Mairal et al. (2010) and Feng et al. (2013) assumed that the dictionary (or basis) is uniformly bounded. Here, we prove that such a condition naturally holds in our case.

Corollary 1 (Uniform bound and Lipschitz of the surrogate) *Following the notation in Proposition 3, we have for all $t > 0$,*

1. $\tilde{\ell}\,(z_t, L_t, \boldsymbol{r}_t, \boldsymbol{e}_t)$ (2.6) *and* $\ell\,(z_t, L_t)$ (2.9) *are both uniformly bounded.*
2. *The surrogate function, i.e., $g_t(L)$ defined in* (3.1) *is uniformly bounded over $\mathcal{L}$.*
3. *Moreover, $g_t(L)$ is uniformly Lipschitz over the compact set $\mathcal{L}$.*

Stage II We next show that the positive stochastic process $\{g_t(L_t)\}_{t=1}^{\infty}$ converges almost surely. To establish the convergence, we verify that $\{g_t(L_t)\}_{t=1}^{\infty}$ is a quasi-martingale (Bottou 1998) that converges almost surely. To this end, we illustrate that the expectation of the discrepancy of $g_{t+1}(L_{t+1})$ and $g_t(L_t)$ can be upper bounded by a family of functions $\ell(\cdot, L)$ indexed by $L \in \mathcal{L}$. Then we show that the family of the functions is P-Donsker (Vaart 2000), the summands of which concentrate around its expectation within an $O(1/\sqrt{n})$ ball almost surely. Therefore, we conclude that $\{g_t(L_t)\}_{t=1}^{\infty}$ is a quasi-martingale and converges almost surely.

Proposition 4 *Let $L \in \mathcal{L}$ and denote the minimizer of $\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e})$ as:*

$$\{\boldsymbol{r}^*, \boldsymbol{e}^*\} = \mathop{\arg\min}_{\boldsymbol{r}, \boldsymbol{e}, \|\boldsymbol{r}\|_2 \leq 1} \frac{1}{2}\|z - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}).$$

Then, the function $\ell(z, L)$ defined in Problem (2.9) is continuously differentiable and

$$\nabla_L \ell(z, L) = (L\boldsymbol{r}^* + \boldsymbol{e}^* - z)\boldsymbol{r}^{*\top}.$$

Furthermore, $\ell(z, \cdot)$ is uniformly Lipschitz over the compact set $\mathcal{L}$.

Proof The gradient of $\ell(z, \cdot)$ follows from Lemma 2. Since each term of $\nabla_L \ell(z, L)$ is uniformly bounded, we conclude the uniform Lipschitz property of $\ell(z, L)$ w.r.t. L.

Corollary 2 (Uniform bound and Lipschitz of the empirical loss) *Let $f_t(L)$ be the empirical loss function defined in (2.8). Then $f_t(L)$ is uniformly bounded and Lipschitz over the compact set $\mathcal{L}$.*

Corollary 3 (P-Donsker of $\ell(z, L)$) *The set of measurable functions $\{\ell(z, L),\ L \in \mathcal{L}\}$ is P-Donsker (see definition in Lemma 1).*

Proposition 5 (Concentration of the empirical loss) *Let $f_t(L)$ and $f(L)$ be the empirical and expected loss functions we defined in (2.8) and (4.1). Then we have*

$$\mathbb{E}[\sqrt{t}\, \| f_t - f \|_\infty] = O(1).$$

Proof Since $\ell(z, L)$ is uniformly upper bounded (Corollary 1) and is always non-negative, its square is uniformly upper bounded, hence its expectation. Together with Corollary 3, Lemma 1 applies.

Theorem 2 (Convergence of the surrogate) *The sequence $\{g_t(L_t)\}_{t=1}^\infty$ we defined in (3.1) converges almost surely, where $\{L_t\}_{t=1}^\infty$ is the solution produced by Algorithm 1. Moreover, the infinite summation $\sum_{t=1}^\infty |\mathbb{E}[g_{t+1}(L_{t+1}) - g_t(L_t) \mid \mathcal{F}_t]|$ is bounded almost surely.*

Proof The theorem follows by showing that the sequence of $\{g_t(L_t)\}_{t=1}^\infty$ is a quasi-martingale, and hence converges almost surely. To see this, we note that for any $t > 0$, the expectation of the difference $g_{t+1}(L_{t+1}) - g_t(L_t)$ conditioned on the past information $\mathcal{F}_t$ is bounded by $\sup_L (f(L) - f_t(L))/(t+1)$, which is of order $O(1/(\sqrt{t}(t+1)))$ due to Proposition 5. Hence, Lemma 3 applies.

Stage III Now we prove that the sequence of the empirical loss function, $\{f_t(L_t)\}_{t=1}^\infty$ defined in (2.8) converges almost surely to the same limit of its surrogate $\{g_t(L_t)\}_{t=1}^\infty$. According to the central limit theorem, we assert that $f_t(L_t)$ also converges almost surely to the expected loss $f(L_t)$ defined in (4.1), implying that $g_t(L_t)$ and $f(L_t)$ converge to the same limit almost surely.

We first establish the numerical convergence of the basis sequence $\{L_t\}_{t=1}^\infty$, based on which we show the convergence of $\{f_t(L_t)\}_{t=1}^\infty$ by applying Lemma 4.

Proposition 6 (Numerical convergence of the basis component) *Let $\{L_t\}_{t=1}^\infty$ be the basis sequence produced by the Algorithm 1. Then, for any $t > 0$, we have*

$$\| L_{t+1} - L_t \|_F = O\left(\frac{1}{t}\right). \tag{4.2}$$

Theorem 3 (Convergence of the empirical and expected loss) *Let $\{f(L_t)\}_{t=1}^\infty$ be the sequence of the expected loss where $\{L_t\}_{t=1}^\infty$ is the sequence of the solutions produced by the Algorithm 1. Then, we have*

1. *The sequence of the empirical loss $\{f_t(L_t)\}_{t=1}^{\infty}$ converges almost surely to the same limit of the surrogate.*
2. *The sequence of the expected loss $\{f(L_t)\}_{t=1}^{\infty}$ converges almost surely to the same limit of the surrogate.*

Proof Let $b_t = g_t(L_t) - f_t(L_t)$. We show that infinite series $\sum_{t=1}^{\infty} b_t/(t+1)$ is bounded by applying the central limit theorem to $f(L_t) - f_t(L_t)$ and the result of Theorem 2. We further prove that $|b_{t+1} - b_t|$ can be bounded by $O(1/t)$, due to the uniform boundedness and Lipschitz of $g_t(L_t)$, $f_t(L_t)$ and $\ell(z_t, L_t)$. According to Lemma 4, we conclude the convergence of $\{b_t\}_{t=1}^{\infty}$ to zero. Hence the first claim. The second claim follows immediately owing to the central limit theorem.

Final stage According to Claim 2 of Theorem 3 and the fact that $\mathbf{0}$ belongs to the subgradient of $g_t(L)$ evaluated at $L = L_t$, we are to show the gradient of $f(L)$ taking at L_t vanishes as t tends to infinity, which establishes Theorem 1. To this end, we note that since $\{L_t\}_{t=1}^{\infty}$ is uniformly bounded, the non-differentiable term $\frac{1}{2t}\|L\|_{2,\infty}^2$ vanishes as t goes to infinity, implying the differentiability of $g_\infty(L_\infty)$, i.e. $\nabla g_\infty(L_\infty) = \mathbf{0}$. On the other hand, we show that the gradient of $f(L)$ and that of $g_t(L)$ are always Lipschitz on the compact set $\mathcal{L}$, implying the existence of their second order derivative even when $t \to \infty$. Thus, by taking a first order Taylor expansion and let t go to infinity, we establish the main theorem.

5 Connection to matrix completion

While we mainly focus on the matrix decomposition problem, our method can be extended to the matrix completion (MC) problem (Cai et al. 2010; Candès and Recht 2009) with max-norm regularization (Cai and Zhou 2013, 2016)—another popular topic in machine learning and signal processing. We focus on the max-norm regularized MC problem with squared Frobenius loss widely considered in the literature, which can be described as follows:

$$\min_{X} \frac{1}{2}\|\mathcal{P}_\Omega(Z - X)\|_F^2 + \frac{\lambda}{2}\|X\|_{\max}^2,$$

where Ω is the set of indices of observed entries in Z and $\mathcal{P}_\Omega(M)$ is the orthogonal projection onto the span of matrices vanishing outside of Ω so that the (i,j)th entry of $\mathcal{P}_\Omega(M)$ is equal to M_{ij} if $(i,j) \in \Omega$ and zero otherwise. Interestingly, the max-norm regularized MC problem can be cast into our framework. To see this, let us introduce an auxiliary matrix M, with $M_{ij} = c > 0$ if $(i,j) \in \Omega$ and $M_{ij} = 1/c$ otherwise. The reformulated MC problem,

$$\min_{X,E} \frac{1}{2}\|Z - X - E\|_F^2 + \frac{\lambda}{2}\|X\|_{\max}^2 + \|M \circ E\|_1, \tag{5.1}$$

where "$\circ$" denotes the entry-wise product, is similar to our MRMD formulation (1.1). And it is easy to show that when c tends to infinity, the reformulated problem converges to the original MC problem.

Online implementation We now derive a stochastic implementation for the max-norm regularized MC problem. Note that the only difference between the Problem (5.1) and Problem (1.1) is the ℓ_1 regularization on E, which results a new penalty on $\boldsymbol{e}$ for $\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e})$ (which is originally defined in (2.6)):

$$\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e}) = \frac{1}{2}\|z - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \|\boldsymbol{m} \circ \boldsymbol{e}\|_1 . \tag{5.2}$$

Here, $\boldsymbol{m}$ is a column of the matrix M in (5.1). According to the definition of M, $\boldsymbol{m}$ is a vector with element value being either c or $1/c$. Let us define two support sets as follows:

$$\Omega_1 \overset{\text{def}}{=} \{i \mid m_i = c, 1 \leq i \leq p\},$$
$$\Omega_2 \overset{\text{def}}{=} \{i \mid m_i = 1/c, 1 \leq i \leq p\},$$

where m_i is the ith element of vector $\boldsymbol{m}$. In this way, the newly defined $\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e})$ can be written as

$$\begin{aligned}\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e}) = &\left(\frac{1}{2}\left\|z_{\Omega_1} - (L\boldsymbol{r})_{\Omega_1} - \boldsymbol{e}_{\Omega_1}\right\|_2^2 + c\left\|\boldsymbol{e}_{\Omega_1}\right\|_1\right) \\ &+ \left(\frac{1}{2}\left\|z_{\Omega_2} - (L\boldsymbol{r})_{\Omega_2} - \boldsymbol{e}_{\Omega_2}\right\|_2^2 + \frac{1}{c}\left\|\boldsymbol{e}_{\Omega_2}\right\|_1\right).\end{aligned} \tag{5.3}$$

Notably, as Ω_1 and Ω_2 are disjoint, given z, L and $\boldsymbol{r}$, the variable $\boldsymbol{e}$ in (5.3) can be optimized by soft-thresholding in a separate manner:

$$\boldsymbol{e}_{\Omega_1} = \mathcal{S}_c\left[z_{\Omega_1} - (L\boldsymbol{r})_{\Omega_1}\right], \quad \boldsymbol{e}_{\Omega_2} = \mathcal{S}_{1/c}\left[z_{\Omega_2} - (L\boldsymbol{r})_{\Omega_2}\right]. \tag{5.4}$$

Hence, we obtain Algorithm 5 for the online max-norm regularized matrix completion (OMRMC) problem. The update principle for $\boldsymbol{r}$ is the same as we described in Algorithm 3 and that for $\boldsymbol{e}$ is given by (5.4). Note that we can use Algorithm 4 to update L as usual.

ℓ_∞-norm constrained variant In some matrix completion applications, one may have to take another ℓ_∞-norm constraint into account, i.e.,

$$\|X\|_\infty \leq \tau, \text{ for some } \tau > 0. \tag{5.5}$$

For example, the rating value of the Netflix dataset is not greater than 5. In the 1-bit setting, the entries of a matrix can either be 1 or -1 (Davenport et al. 2014). Other examples can be found in, e.g., Klopp (2014). Interestingly, Algorithm 5 can be adjusted to such a constraint.

To see this, we observe that the constraint $\|X\|_\infty \leq \tau$ amounts to restricting

$$\left|x_{ij}\right| \leq \tau$$

for all entries x_{ij} of X. Due to the matrix factorization $X = LR^\top$, we know that it requires

$$\left|\boldsymbol{l}(i)\boldsymbol{r}(j)^\top\right| \leq \tau, \ \forall\, i \in [p], \ \forall\, j \in [n], \tag{5.6}$$

where we recall that $\boldsymbol{l}(i)$ and $\boldsymbol{r}(j)$ are the ith row of L and the jth row of R, respectively. Proposition 1 already ensures

$$\|\boldsymbol{r}(j)\|_2 \leq 1, \ \forall\, j \in [n].$$

Since $\left|\boldsymbol{l}(i)\boldsymbol{r}(j)^\top\right| \leq \|\boldsymbol{l}(i)\|_2 \cdot \|\boldsymbol{r}(j)\|_2$, we obtain a *sufficient* condition for (5.6):

$$\|\boldsymbol{l}(i)\|_2 \leq \tau, \ \forall\, i \in [n].$$

That is,

$$\|L\|_{2,\infty} \leq \tau,$$

which can easily be fulfilled by an orthogonal projection onto the ℓ_2 ball with radius τ, i.e., if $\|L_t\|_{2,\infty} > \tau$, we set $L_t \leftarrow \frac{\tau}{\|L_t\|_{2,\infty}} L_t$.

Algorithm 5 Online Max-Norm Regularized Matrix Completion

```
Require: $Z \in \mathbb{R}^{p \times n}$ (observed samples), parameters $\lambda_1$ and $\lambda_2$, $L_0 \in \mathbb{R}^{p \times d}$ (initial basis), zero matrices
    $A_0 \in \mathbb{R}^{d \times d}$ and $B_0 \in \mathbb{R}^{p \times d}$
Ensure: optimal basis $L_t$
for $t = 1$ to $n$ do
    Access the $t$th sample $z_t$.
    Compute the coefficient and noise:
        $\{\boldsymbol{r}_t, \boldsymbol{e}_t\} = \underset{\boldsymbol{r}, \boldsymbol{e}, \|\boldsymbol{r}\|_2^2 \le 1}{\arg\min} \; \tilde{\ell}(z_t, L_{t-1}, \boldsymbol{r}, \boldsymbol{e})$
        $= \underset{\boldsymbol{r}, \boldsymbol{e}, \|\boldsymbol{r}\|_2^2 \le 1}{\arg\min} \left( \frac{1}{2} \|z_t - L_{t-1}\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \|\boldsymbol{m}_t \circ \boldsymbol{e}\|_1 \right).$
    Compute the accumulation matrices $A_t$ and $B_t$:
        $A_t \leftarrow A_{t-1} + \boldsymbol{r}_t \boldsymbol{r}_t^\top,$
        $B_t \leftarrow B_{t-1} + (z_t - \boldsymbol{e}_t)\, \boldsymbol{r}_t^\top.$
    Compute the basis $L_t$ by optimizing the surrogate function (3.1):
        $L_t = \underset{L}{\arg\min} \frac{1}{t} \sum_{i=1}^{t} \tilde{\ell}(z_i, L, \boldsymbol{r}_i, \boldsymbol{e}_i) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2$
        $= \underset{L}{\arg\min} \frac{1}{t} \sum_{i=1}^{t} \left( \frac{1}{2} \|z_i - L\boldsymbol{r}_i - \boldsymbol{e}_i\|_2^2 + \|\boldsymbol{m}_i \circ \boldsymbol{e}_i\|_1 \right) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2$
        $= \underset{L}{\arg\min} \frac{1}{t} \sum_{i=1}^{t} \left( \frac{1}{2} \|z_i - L\boldsymbol{r}_i - \boldsymbol{e}_i\|_2^2 \right) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2$
        $= \underset{L}{\arg\min} \frac{1}{t} \left( \frac{1}{2} \operatorname{Tr}\left( L^\top L A_t \right) - \operatorname{Tr}\left( L^\top B_t \right) \right) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2.$
end for
```

Other types of loss functions We in this paper emphasize on the squared Frobenius loss for the max-norm regularized problems. There is also solid theoretical analysis for other formulations, e.g., logistic regression and probit regression (Cai and Zhou 2013). Unfortunately, it seems that one cannot trivially extend the proposed online algorithms to a general loss function. To be more precise, for Frobenius (or ℓ_2) loss, we are guaranteed with a nice property that minimizing the surrogate (3.8) is equivalent to solving (3.10), for which only $O(pd)$ memory is needed. For general models, such a property does not hold and we conjecture that more technique is needed to find a good approximation to (3.8).

6 Experiments

In this section, we report numerical results on synthetic data to demonstrate the effectiveness and robustness of our online max-norm regularized matrix decomposition (OMRMD) algorithm. Some experimental settings are used throughout this section, as elaborated below.

Data generation The simulation data are generated by following a similar procedure in Candès et al. (2011). The clean data matrix X is produced by $X = UV^\top$, where $U \in \mathbb{R}^{p \times d}$ and $V \in \mathbb{R}^{n \times d}$. The entries of U and V are i.i.d. sampled from the normal distribution $\mathcal{N}(0, 1)$. We choose sparse corruption in the experiments, and introduce a parameter ρ to control the sparsity of the corruption matrix E, *i.e.*, a ρ-fraction of the entries are non-zero whose locations are uniformly sampled and the magnitude follows a uniform distribution over $[-1000, 1000]$. Finally, the observation matrix Z is produced by $Z = X + E$.

Baselines We mainly compare with two methods: Principal Component Pursuit (PCP) and online robust PCA (OR-PCA). PCP is the state-of-the-art batch method for subspace recovery, which was presented as a robust formulation of PCA in Candès et al. (2011). OR-PCA is an online implementation of PCP,[2] which also achieves state-of-the-art performance over the online subspace recovery algorithms. Sometimes, to show the robustness, we will also report the results of online PCA (Artač et al. 2002), which incrementally learns the principal components without taking the noise into account.

Evaluation metric Our goal is to estimate the correct subspace for the underlying data. Here, we evaluate the fitness of our estimated subspace basis L and the ground truth basis U by the Expressed Variance (EV) (Xu et al. 2010):

$$\mathrm{EV}(U, L) \overset{\text{def}}{=} \frac{\mathrm{Tr}(L^\top U U^\top L)}{\mathrm{Tr}(U U^\top)}. \tag{6.1}$$

The values of EV range in [0, 1] and a higher value indicates a more accurate recovery.

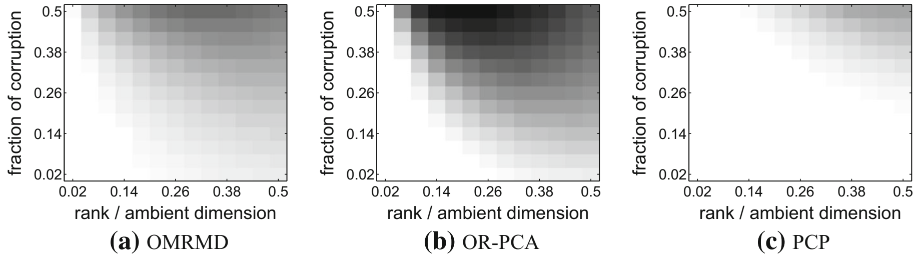
Other settings Throughout the experiments, we set the ambient dimension $p = 400$, the total number of samples $n = 5000$ and pick the value of d as the true rank unless otherwise specified. We fix the tunable parameter $\lambda_1 = \lambda_2 = 1/\sqrt{p}$, and use default parameters for all baselines we compare with. Each experiment is repeated 10 times and we report the averaged EV as the result.

6.1 Robustness

We first study the robustness of OMRMD, measured by the EV value of its output after accessing the last sample, and compare it to the nuclear norm based OR-PCA and the batch algorithm PCP. In order to make a detailed examination, we vary the true rank from $0.02p$ to $0.5p$, with a step size $0.04p$, and the corruption fraction ρ from 0.02 to 0.5, with a step size 0.04.

The general results are illustrated in Fig. 1 where a brighter color means a higher EV (hence better performance). We observe that for easy tasks (i.e., few corruption and low rank case), both OMRMD and OR-PCA perform comparably. However, for more difficult cases, OMRMD outperforms OR-PCA. In order to further investigate this phenomenon, we plot the EV curve against the fraction of corruption under a given matrix rank. In particular, we group the results into two parts, one with relatively low rank (Fig. 2) and the other with middle level of rank (Fig. 3). Figure 2 indicates that when manipulating a low-rank matrix, OR-PCA works as well as OMRMD under a low level of noise. For instance, the EV produced by OR-PCA is as close as that of OMRMD for rank less than 40 and ρ no more than 0.26. However, when the rank becomes larger, OR-PCA degrades quickly compared to OMRMD.

[2] Strictly speaking, OR-PCA is an online version of stable PCP (Zhou et al. 2010).



(a) OMRMD **(b)** OR-PCA **(c)** PCP

Fig. 1 Performance of subspace recovery under different rank and corruption fraction. Brighter color means better performance. As we observed, the max-norm based algorithm OMRMD always performs comparably or better than OR-PCA which is based on nuclear norm formulation. Since PCP is a batch method, it always achieves the best recovery performance

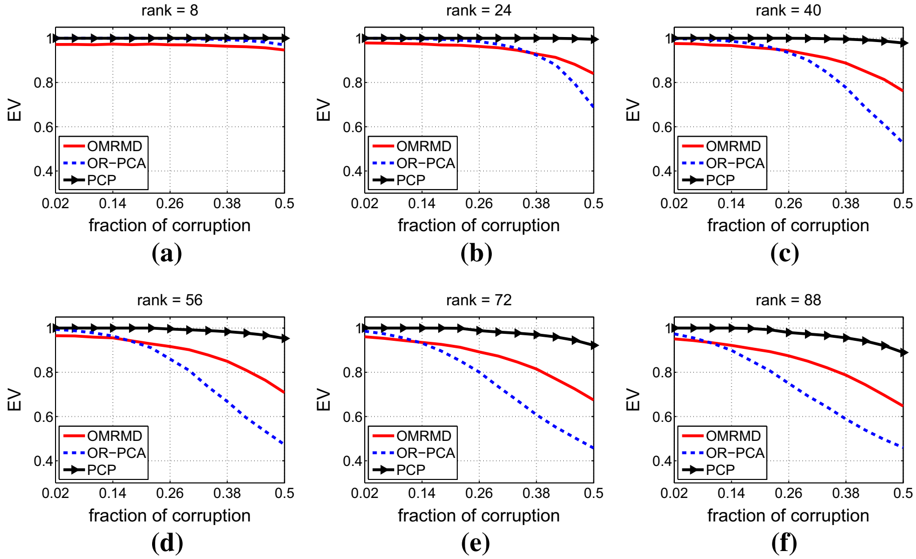
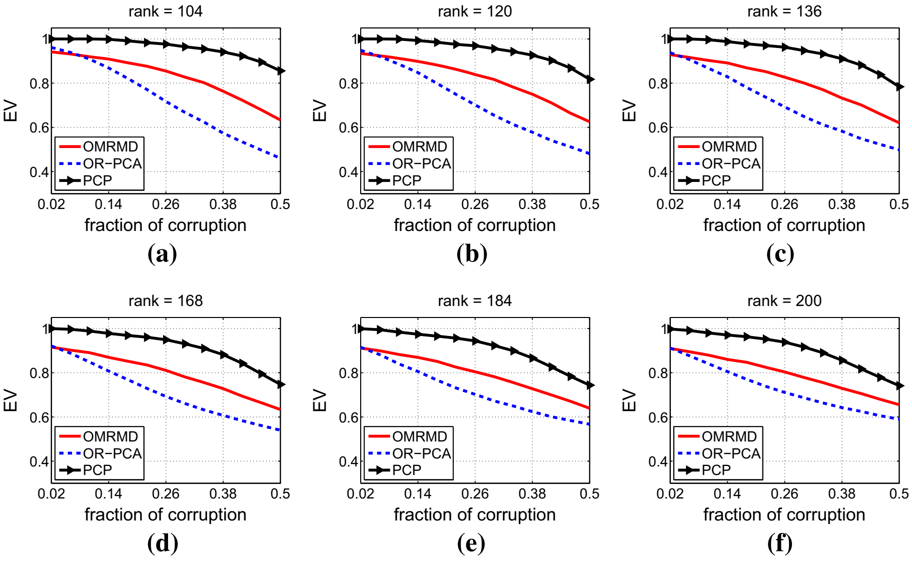


Fig. 2 EV value against corruption fractions when the matrix has a relatively low rank (note that the ambient dimension p is 400). The EV value is computed for the obtained basis after accessing the last sample. When the rank is extremely low (rank $= 8$), OMRMD and OR-PCA works comparably. In other cases, OMRMD is always better than OR-PCA addressing a large fraction of corruption

This is possibly because the max-norm is a tighter approximation to the matrix rank. Since PCP is a batch formulation and accesses all the data in each iteration, it always achieves the best recovery performance.

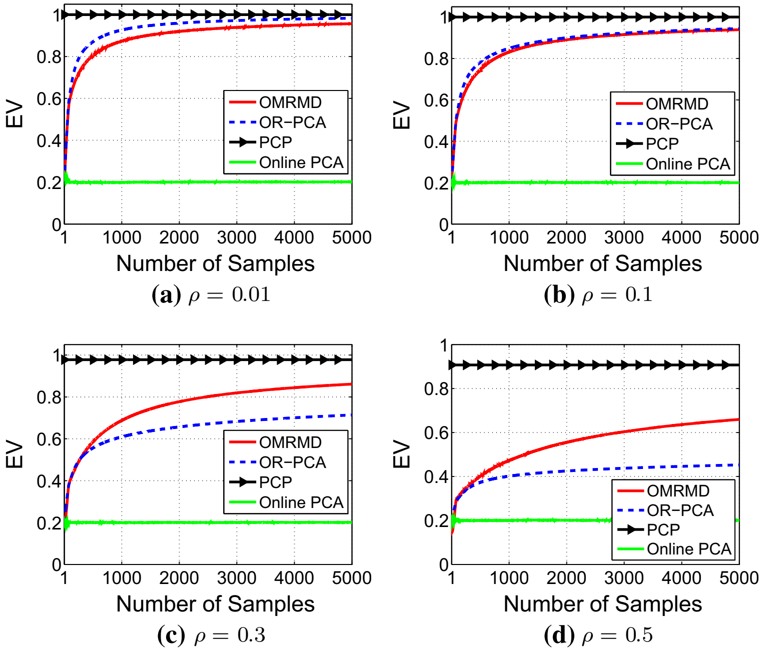
6.2 Convergence rate

We next study the convergence of OMRMD by plotting the EV curve against the number of samples. Besides OR-PCA and PCP, we also add online PCA (Artač et al. 2002) as a baseline algorithm. The results are illustrated in Fig. 4 where we set $p = 400$ and the true rank as 80. As expected, PCP achieves the best performance since it is a batch method and needs to access all the data during optimization. Online PCA degrades significantly even with low corruption (Fig. 4a). OMRMD is comparable to OR-PCA when the corruption is



(a) (b) (c) (d) (e) (f)

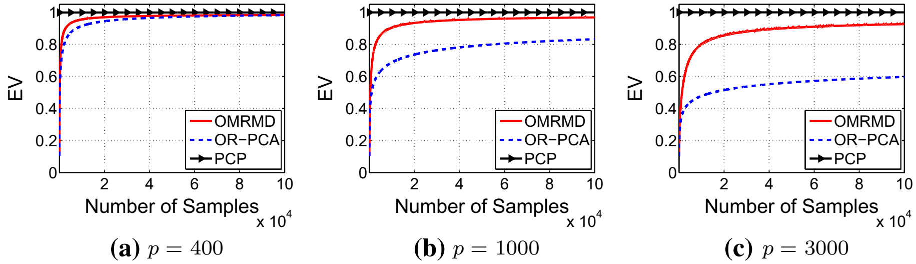
Fig. 3 EV value against corruption fractions when the matrix has a middle level of rank (note that the ambient dimension p is 400). The EV value is computed for the basis after accessing the last sample. In these cases, OR-PCA degrades as soon as the corruption is tuned to be higher than 0.02



(a) $\rho = 0.01$ **(b)** $\rho = 0.1$

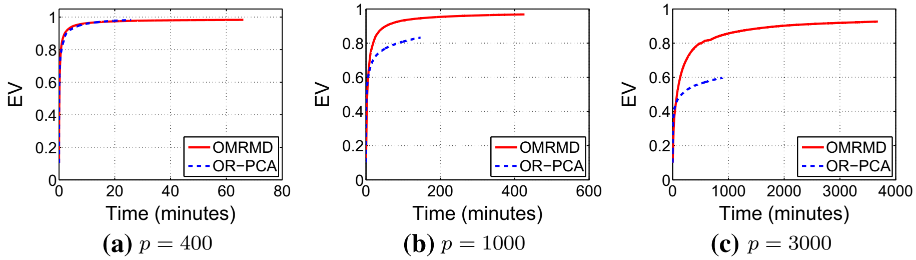
(c) $\rho = 0.3$ **(d)** $\rho = 0.5$

Fig. 4 EV value against number of samples under different corruption fractions. PCP outperforms all the online algorithms before they converge since PCP accesses all the data to estimate the basis. The performance of Online PCA is significantly degraded even when there is little corruption. For hard tasks (ρ equal to 0.3 or higher), we again observe the superiority of the max-norm over the nuclear norm



(a) $p = 400$ **(b)** $p = 1000$ **(c)** $p = 3000$

Fig. 5 EV value against number of samples under different ambient dimensions. The intrinsic dimension $d = 0.1p$ and the corruption fraction $\rho = 0.3$



(a) $p = 400$ **(b)** $p = 1000$ **(c)** $p = 3000$

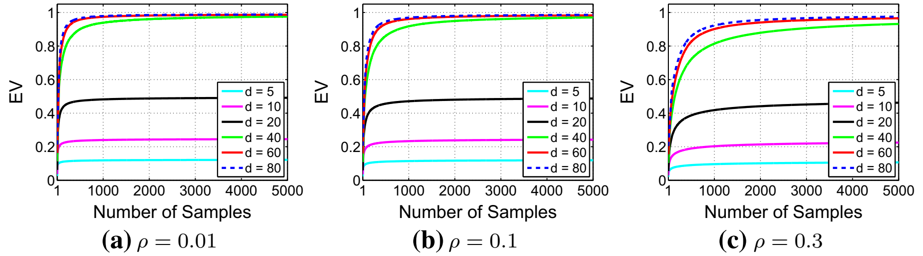
Fig. 6 EV value against time under different ambient dimensions. The intrinsic dimension d is set as $0.1p$ and the corruption fraction ρ equals 0.3

low (Fig. 4a), and converges significantly faster when the data is grossly corrupted (Fig. 4c and 4d). This observation agrees with Fig. 1, and again suggests that in the noisy scenario, max-norm may be a better fit than the nuclear norm.

Indeed, OMRMD converges much faster even in large scale problems. In Fig. 5, we compare the convergence rate of OMRMD and OR-PCA under different ambient dimensions. The rand of the data are set with $0.1p$, indicating a low-rank structure of the underlying data. Again, we assume the rank is known so $d = 0.1p$. The error corruption ρ is fixed to 0.3 – a difficult task for recovery. We observe that for high dimensional cases ($p = 1000$ and $p = 3000$), OMRMD significantly outperforms OR-PCA. For example, in Fig. 5b, OMRMD achieves the EV value of 0.8 only with accessing about 2000 samples, whereas OR-PCA needs to reveal 60, 000 samples to obtain the same accuracy!

6.3 Computational complexity

We note that OMRMD is a little bit inferior to OR-PCA in terms of computation per iteration, as our algorithm may solve a dual problem to optimize $\boldsymbol{r}$ (see Algorithm 3) if the initial solution $\boldsymbol{r}_0$ violates the constraint. We plot the EV curve with respect to the running time in Fig. 6. It shows that, OR-PCA is about 3 times faster than OMRMD when processing a data point. However, we point out here that we emphasize on the convergence rate. That is, given an EV value, how much time the algorithm will take to achieve it. In Fig. 6c, for example, OMRMD takes 50 minutes to achieve the EV value of 0.6, while OR-PCA uses nearly 900 minutes. From Figs. 5 and 6, it is safe to conclude that OMRMD is superior to OR-PCA in terms of convergence rate in the price of a little more computation per sample.



(a) $\rho = 0.01$ **(b)** $\rho = 0.1$ **(c)** $\rho = 0.3$

Fig. 7 Influence of the choice of d. The true matrix rank is 40. We observe that as long as d is no smaller than the true rank, the algorithm always recovers the subspace

6.4 Influence of *d*

Finally, we remark that it is important to pick a sufficiently large value for d. As Burer and Monteiro (2005) suggested, d should be chosen no smaller than the true rank. In the simulation studies, we always pick d as the rank of the underlying data. Here we examine the influence of d in Fig. 7, where we set the ambient dimension $p = 400$, the sample size $n = 5000$ and the true rank is 40. As expected, if the value of d is smaller than the true rank, we have no hope to recover the subspace.

7 Conclusion

In this paper, we have developed an online algorithm for the max-norm regularized matrix decomposition problems. Using the matrix factorization form of the max-norm, we converted the original problem to a constrained one which facilitates an online implementation for solving the batch problem. We have established theoretical guarantees that the sequence of the solutions converges to a stationary point of the expected loss function asymptotically. Moreover, we empirically compared our proposed algorithm with OR-PCA, which is a recently proposed online algorithm for nuclear-norm based matrix decomposition. The simulation results have suggested that the proposed algorithm is more robust than OR-PCA, in particular for hard tasks (i.e., when a large fraction of entries are corrupted). We also have investigated the convergence rate for both OMRMD and OR-PCA, and have shown that OMRMD converges much faster than OR-PCA even in large-scale problems. When acquiring sufficient samples, we observed that our algorithm converges to the batch method PCP, which is a state-of-the-art formulation for subspace recovery. Our experiments, to an extent, suggest that the max-norm might be a tighter relaxation of the rank function compared to the nuclear norm.

Acknowledgements J. Shen and P. Li are partially supported by NSF-Bigdata-1419210, NSF-III-1360971, ONR-N00014-13-1-0764 and AFOSR-FA9550-13-1-0137.

8 Appendix: Proof details

8.1 Proof for Proposition 1

Proof Let us denote $k = \|R\|_{2,\infty}$. We presume that k is positive. Otherwise, the low-rank component X we aim to recover is a zero matrix, which is of little interest. Now we construct two auxiliary variables $\bar{L} = kL \in \mathbb{R}^{p\times d}$ and $\bar{R} = \frac{1}{k}R \in \mathbb{R}^{n\times d}$. Replacing L and R with $\frac{1}{k}\bar{L}$ and $k\bar{R}$ in (2.2) respectively, we have:

$$\min_{\bar{L},\bar{R},E} \frac{1}{2}\left\| Z - \left(\frac{1}{k}\bar{L}\right)\left(k\bar{R}\right)^\top - E \right\|_F^2 + \frac{\lambda_1}{2}\left\|\frac{1}{k}\bar{L}\right\|_{2,\infty}^2 \left\|k\bar{R}\right\|_{2,\infty}^2 + \lambda_2 h(E).$$

That is, we are to solve

$$\min_{\bar{L},\bar{R},E} \frac{1}{2}\left\| Z - \bar{L}\bar{R}^\top - E \right\|_F^2 + \frac{\lambda_1}{2}\left\|\bar{L}\right\|_{2,\infty}^2 \left\|\bar{R}\right\|_{2,\infty}^2 + \lambda_2 h(E).$$

The fact that $\bar{R} = \frac{1}{k}R$ and k is the maximum of the ℓ_2 row norm of R implies $\left\|\bar{R}\right\|_{2,\infty} = 1$. Therefore, we can reformulate our MRMD problem as a constrained program:

$$\min_{\bar{L},\bar{R},E} \frac{1}{2}\left\| Z - \bar{L}\bar{R}^\top - E \right\|_F^2 + \frac{\lambda_1}{2}\left\|\bar{L}\right\|_{2,\infty}^2 + \lambda_2 h(E), \quad \text{s. t. } \left\|\bar{R}\right\|_{2,\infty}^2 = 1.$$

To see why the above program is equivalent to (2.3), we only need to show that each optimal solutions (L^*, R^*, E^*) of (2.3) must satisfy $\|R^*\|_{2,\infty}^2 = 1$. Suppose that $k = \|R^*\|_{2,\infty} < 1$. Let $L' = kL^*$ and $R' = \frac{1}{k}R^*$. Obviously, (L', R', E^*) are still feasible. However, the objective value becomes

$$\begin{aligned}
&\frac{1}{2}\left\| Z - L'R'^\top - E^* \right\|_F^2 + \frac{\lambda_1}{2}\left\|L'\right\|_{2,\infty}^2 + \lambda_2 h(E^*) \\
&\quad = \frac{1}{2}\left\| Z - L^*R^{*\top} - E^* \right\|_F^2 + \frac{\lambda_1}{2}\cdot k^2\left\|L^*\right\|_{2,\infty}^2 + \lambda_2 h(E^*) \\
&\quad < \frac{1}{2}\left\| Z - L^*R^{*\top} - E^* \right\|_F^2 + \frac{\lambda_1}{2}\left\|L^*\right\|_{2,\infty}^2 + \lambda_2 h(E^*),
\end{aligned}$$

which contradicts the assumption that (L^*, R^*, E^*) is the optimal solution. Thus we complete the proof.

8.2 Proof for Stage I

First we prove that all the stochastic variables are uniformly bounded.

Proposition 7 *Let $\boldsymbol{r}_t$, $\boldsymbol{e}_t$ and L_t be the optimal solutions produced by Algorithm 1. Then,*

1. *The optimal solutions $\boldsymbol{r}_t$ and $\boldsymbol{e}_t$ are uniformly bounded.*
2. *The matrices $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$ are uniformly bounded.*
3. *There exists a compact set $\mathcal{L}$, such that for all L_t produced by Algorithm 1, $L_t \in \mathcal{L}$. Namely, there exists a positive constant $L_{\max}$ that is uniform over t, such that for all $t > 0$,*

$$\|L_t\|_F \leq L_{\max}.$$

Proof Note that for each $t > 0$, $\|\boldsymbol{r}_t\|_2 \leq 1$. Thus $\boldsymbol{r}_t$ is uniformly bounded. Let us consider the optimization problem (3.2). As the trivial solution $\boldsymbol{r}_t = \boldsymbol{0}$ and $\boldsymbol{e}_t = \boldsymbol{0}$ are feasible, we have

$$\tilde{\ell}(\boldsymbol{z}_t, L_{t-1}, \boldsymbol{0}, \boldsymbol{0}) = \frac{1}{2}\|\boldsymbol{z}_t\|_2^2.$$

Therefore, the optimal solution should satisfy:

$$\frac{1}{2}\|\boldsymbol{z}_t - L_{t-1}\boldsymbol{r}_t - \boldsymbol{e}_t\|_2^2 + \lambda_2\|\boldsymbol{e}_t\|_1 \leq \frac{1}{2}\|\boldsymbol{z}_t\|_2^2,$$

which implies

$$\|\boldsymbol{e}_t\|_1 \le \frac{1}{2\lambda_2}\|\boldsymbol{z}_t\|_2^2 .$$

Since $\boldsymbol{z}_t$ is uniformly bounded (Assumption $(A1)$), $\boldsymbol{e}_t$ is uniformly bounded.

To examine the uniform bound for $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$, note that

$$\frac{1}{t}A_t = \frac{1}{t}\sum_{i=1}^{t}\boldsymbol{r}_i\boldsymbol{r}_i^\top, \quad \frac{1}{t}B_t = \frac{1}{t}\sum_{i=1}^{t}(\boldsymbol{z}_i - \boldsymbol{e}_i)\,\boldsymbol{r}_i^\top .$$

Since for each i, $\boldsymbol{r}_i$, $\boldsymbol{e}_i$ and $\boldsymbol{z}_i$ are uniformly bounded, $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$ are uniformly bounded.

Based on Claim 1 and Claim 2, we prove that L_t can be uniformly bounded. First let us denote $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$ by $\widetilde{A}_t$ and $\widetilde{B}_t$, respectively.

Step 1 According to Claim 2, there exist constants a_1 and b that are uniform over t, such that

$$\left\|\widetilde{A}_t\right\|_F \le a_1, \quad \left\|\widetilde{B}_t\right\|_F \le b.$$

On the other hand, from Assumption $(A2)$, the eigenvalues of $\widetilde{A}_t$ is lower bounded by a positive constant β_1 that is uniform over t, implying the trace norm (sum of the singular values) of $\widetilde{A}_t$ is uniformly lower bounded by a positive constant. As all norms are equivalent, we can show that

$$\left\|\widetilde{A}_t\right\|_F \ge a_0 > 0,$$

where a_0 is a positive constant which is uniform over t.

Recall that L_t is the optimal basis for (3.10). Thus, the subgradient of the objective function taken at L_t should contain zero, that is,

$$L_t\widetilde{A}_t - \widetilde{B}_t + \frac{\lambda_1}{t}U_t = 0,$$

where U_t is the subgradient of $\frac{1}{2}\|L_t\|_{2,\infty}^2$ produced by (3.11). Note that, as all of the eigenvalues of $\widetilde{A}_t$ are lower bounded by a positive constant, $\widetilde{A}_t$ is invertible. Thus,

$$L_t = \left(\widetilde{B}_t - \frac{\lambda_1}{t}U_t\right)\widetilde{A}_t^{-1},$$

where $\widetilde{A}_t^{-1}$ is the inverse of $\widetilde{A}_t$.

Now we derive the bound for L_t:

$$\begin{aligned}
\|L_t\|_F &= \left\|\left(\widetilde{B}_t - \frac{\lambda_1}{t}U_t\right)\widetilde{A}_t^{-1}\right\|_F \\
&\le \left\|\widetilde{B}_t - \frac{\lambda_1}{t}U_t\|_F \cdot \|\widetilde{A}_t^{-1}\right\|_F \\
&\le \left(\left\|\widetilde{B}_t\right\|_F + \frac{\lambda_1}{t}\|U_t\|_F\right)\left\|\widetilde{A}_t^{-1}\right\|_F \\
&= \left\|\widetilde{A}_t^{-1}\right\|_F\left\|\widetilde{B}_t\right\|_F + \frac{\lambda_1}{t}\left\|\widetilde{A}_t^{-1}\right\|_F\|U_t\|_F \\
&\le \left\|\widetilde{A}_t^{-1}\right\|_F\left\|\widetilde{B}_t\right\|_F + \frac{\lambda_1}{t}\left\|\widetilde{A}_t^{-1}\right\|_F\|L_t\|_F .
\end{aligned}$$

It follows that

$$\left(1 - \frac{\lambda_1}{t} \left\| \widetilde{A}_t^{-1} \right\|_F\right) \|L_t\|_F \leq \left\| \widetilde{A}_t^{-1} \right\|_F \left\| \widetilde{B}_t \right\|_F .$$

As all of the eigenvalues of $\widetilde{A}_t$ are uniformly lower bounded, those of $\widetilde{A}_t^{-1}$ are uniformly upper bounded. Thus the trace norm of $\widetilde{A}_t^{-1}$ are uniformly upper bounded. As all norms are equivalent, $\|\widetilde{A}_t^{-1}\|_F$ is also uniformly upper bounded by a constant, say a_2. Thus,

$$\left(1 - \frac{\lambda_1}{t} a_2\right) \|L_t\|_F \leq \left(1 - \frac{\lambda_1}{t} \left\| \widetilde{A}_t^{-1} \right\|_F\right) \|L_t\|_F \leq \left\| \widetilde{A}_t^{-1} \right\|_F \left\| \widetilde{B}_t \right\|_F \leq a_2 b$$

Particularly, let

$$t_0 = \min_t \{t \geq 2\lambda_1 a_2, t \text{ is an integer}\} .$$

Then, for all $t \geq t_0$,

$$\|L_t\|_F \leq 2a_2 b. \tag{8.1}$$

Step 2 Let us consider a uniform bound for L_t, with $0 < t < t_0$. Recall that L_t is the minimizer for $g_t(L)$, that is

$$\begin{aligned} L_t &= \arg\min_L g_t(L) \\ &= \arg\min_L \frac{1}{t} \sum_{i=1}^{t} \left(\frac{1}{2} \|\boldsymbol{z}_i - L\boldsymbol{r}_i - \boldsymbol{e}_i\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}_i)\right) + \frac{\lambda_1}{2t} \|L\|_{2,\infty}^2 \\ &= \arg\min_L \sum_{i=1}^{t} \frac{1}{2} \|\boldsymbol{z}_i - L\boldsymbol{r}_i - \boldsymbol{e}_i\|_2^2 + \frac{\lambda_1}{2} \|L\|_{2,\infty}^2 \\ &\overset{\text{def}}{=} \arg\min_L \tilde{g}_t(L). \end{aligned}$$

Consider a trivial but feasible solution with $L = 0$,

$$\tilde{g}_t(0) = \sum_{i=1}^{t} \frac{1}{2} \|\boldsymbol{z}_i - \boldsymbol{e}_i\|_2^2 .$$

The inequality

$$\tilde{g}_t(L_t) \leq \tilde{g}_t(0)$$

implies

$$\|L_t\|_{2,\infty}^2 \leq \frac{1}{\lambda_1} \sum_{i=1}^{t} \|\boldsymbol{z}_i - \boldsymbol{e}_i\|_2^2 .$$

Since

$$\|L_t\|_F^2 \leq p \|L_t\|_{2,\infty}^2 \leq \frac{p}{\lambda_1} \sum_{i=1}^{t} \|\boldsymbol{z}_i - \boldsymbol{e}_i\|_2^2 ,$$

we have

$$\|L_t\|_F \leq \sqrt{\frac{p}{\lambda_1} \sum_{i=1}^{t} \|\boldsymbol{z}_i - \boldsymbol{e}_i\|_2^2}.$$

For all $0 < t < t_0$,

$$\|L_t\|_F \leq \sqrt{\frac{p}{\lambda_1}\sum_{i=1}^{t}\|z_i - e_i\|_2^2} \leq \sqrt{\frac{p}{\lambda_1}\sum_{i=1}^{t_0}\|z_i - e_i\|_2^2}. \tag{8.2}$$

Note that each term, particularly t_0, can be uniformly upper bounded, thus $\sqrt{\frac{p}{\lambda_1}\sum_{i=1}^{t_0}\|z_i - e_i\|_2^2}$ can also be uniformly upper bounded. Namely, for all $0 < t < t_0$, L_t is also uniformly upper bounded.

Step 3 Now let us define

$$L_{\max} = \max\left\{2a_2 b, \sqrt{\frac{p}{\lambda_1}\sum_{i=1}^{t_0}\|z_i - e_i\|_2^2}\right\}.$$

Then, for all $t > 0$,

$$\|L_t\|_F \leq L_{\max}.$$

Remark 4 We remark some critical points in the third claim of Proposition 3. All the constants, a_0, a_1, a_2 and b are independent from t, making them uniformly bounded. Also, t_0 is a constant that is uniform over t. Thus, L_t can be uniformly bounded.

Corollary 4 *Let $\boldsymbol{r}_t$, $\boldsymbol{e}_t$ and L_t be the optimal solutions produced by Algorithm 1. We show some uniform boundedness property here.*

1. *$\tilde{\ell}(z_t, L_t, \boldsymbol{r}_t, \boldsymbol{e}_t)$ defined in (2.6) and $\ell(z_t, L_t)$ defined in (2.9) are both uniformly bounded.*
2. *The surrogate function,* i.e., *$g_t(L_t)$ defined in (3.1) is uniformly bounded.*
3. *Moreover, $g_t(L)$ is uniformly Lipschitz over the compact set $\mathcal{L}$.*

Proof The uniform bound of $\boldsymbol{r}_t$, $\boldsymbol{e}_t$ and z_t, combined with the uniform bound of L_t, implies the uniform boundedness for $\tilde{\ell}(z_t, L_t, \boldsymbol{r}_t, \boldsymbol{e}_t)$ and $\ell(z_t, L_t)$. Thus, $g_t(L_t)$ and $f_t(L_t)$ are also uniformly bounded.

To show that $g_t(L)$ is uniformly Lipschitz, we compute its subgradient at any $L \in \mathcal{L}$:

$$\begin{aligned}\left\|\nabla_L g_t(L)\right\|_F = \left\|\frac{1}{t}(LA_t - B_t) + \frac{\lambda_1}{t}U\right\|_F &\leq \left\|\frac{1}{t}(LA_t - B_t)\right\|_F + \frac{\lambda_1}{t}\|L\|_F \\ &\leq \left\|\frac{1}{t}(LA_t - B_t)\right\|_F + \lambda_1\|L\|_F\end{aligned}$$

where $U \in \partial\frac{1}{2}\|L\|_{2,\infty}$. Since L, $\frac{1}{t}A_t$ and $\frac{1}{t}B_t$ are all uniformly bounded, the subgradient of $g_t(L)$ is uniformly bounded. This implies that $g_t(L)$ is uniformly Lipschitz.

8.3 Proof for Stage II

Lemma 1 (A corollary of Donsker theorem (Vaart 2000)) *Let $F = \{f_\theta : \mathcal{X} \to \mathbb{R}, \theta \in \Theta\}$ be a set of measurable functions indexed by a bounded subset Θ of $\mathbb{R}^d$. Suppose that there exists a constant K such that*

$$\left|f_{\theta_1}(x) - f_{\theta_2}(x)\right| \leq K\|\theta_1 - \theta_2\|_2,$$

for every θ_1 and θ_2 in Θ and x in $\mathcal{X}$. Then, F is P-Donsker. For any f in F, let us define $\mathbb{P}_n f$, $\mathbb{P} f$ and $\mathbb{G}_n f$ as

$$\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^{n} f(X_i), \quad \mathbb{P} f = \mathbb{E}[f(X)], \quad \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P} f).$$

Let us also suppose that for all f, $\mathbb{P} f^2 < \delta^2$ and $\|f\|_\infty < M$ and that the random elements $X_1, X_2, \ldots$ are Borel-measurable. Then, we have

$$\mathbb{E}\,\|\mathbb{G}\|_F = O(1),$$

where $\|\mathbb{G}\|_F = \sup_{f \in F} |\mathbb{G}_n f|$.

Now let us verify that the set of functions $\{\ell(\boldsymbol{z}, L), L \in \mathcal{L}\}$ indexed by L fulfills the hypotheses in the corollary of Donsker Theorem. In particular, we have verified that:

- The index set $\mathcal{L}$ is uniformly bounded (see Proposition 3).
- Each $\ell(\boldsymbol{z}, L)$ can be uniformly bounded (see Corollary 1).
- Any of the functions $\ell(\boldsymbol{z}, L)$ in the family is uniformly Lipschitz (see Proposition 4).

Next, we show that the family of functions $\ell(\boldsymbol{z}, L)$ is uniformly Lipschitz w.r.t. L. We introduce the following lemma as it will be useful for our discussion.

Lemma 2 (Corollary of Theorem 4.1 from *Bonnans and Shapiro (1998)*) *Let $f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R}$. Suppose that for all $\boldsymbol{x} \in \mathbb{R}^p$ the function $f(\boldsymbol{x}, \cdot)$ is differentiable, and that f and $\nabla_{\boldsymbol{u}} f(\boldsymbol{x}, \boldsymbol{u})$ are continuous on $\mathbb{R}^p \times \mathbb{R}^q$. Let $\boldsymbol{v}(\boldsymbol{u})$ be the optimal value function $\boldsymbol{v}(\boldsymbol{u}) = \min_{\boldsymbol{x} \in \mathcal{C}} f(\boldsymbol{x}, \boldsymbol{u})$, where $\mathcal{C}$ is a compact subset of $\mathbb{R}^p$. Then $\boldsymbol{v}(\boldsymbol{u})$ is directionally differentiable. Furthermore, if for $\boldsymbol{u}_0 \in \mathbb{R}^q$, $f(\cdot, \boldsymbol{u}_0)$ has unique minimizer $\boldsymbol{x}_0$ then $\boldsymbol{v}(\boldsymbol{u})$ is differentiable in $\boldsymbol{u}_0$ and $\nabla_{\boldsymbol{u}} \boldsymbol{v}(\boldsymbol{u}_0) = \nabla_{\boldsymbol{u}} f(\boldsymbol{x}_0, \boldsymbol{u}_0)$.*

Proposition 8 *Let $L \in \mathcal{L}$ and denote the minimizer of $\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e})$ defined in (2.9) as:*

$$\{\boldsymbol{r}^*, \boldsymbol{e}^*\} = \underset{\boldsymbol{r}, \boldsymbol{e}, \|\boldsymbol{r}\|_2 \le 1}{\arg\min} \frac{1}{2}\|\boldsymbol{z} - L\boldsymbol{r} - \boldsymbol{e}\|_2^2 + \lambda_2 \tilde{h}(\boldsymbol{e}).$$

Then, the function $\ell(\boldsymbol{z}, L)$ defined in Problem (2.9) is continuously differentiable and

$$\nabla_L \ell(\boldsymbol{z}, L) = (L\boldsymbol{r}^* + \boldsymbol{e}^* - \boldsymbol{z})\boldsymbol{r}^{*\top}.$$

Furthermore, $\ell(\boldsymbol{z}, \cdot)$ is uniformly Lipschitz.

Proof By fixing the variable $\boldsymbol{z}$, the function $\tilde{\ell}$ can be seen as a mapping:

$$\mathbb{R}^{d+p} \times \mathcal{L} \to \mathbb{R}$$
$$([\boldsymbol{r};\ \boldsymbol{e}], L) \mapsto \tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e}).$$

It is easy to show that $\forall [\boldsymbol{r};\ \boldsymbol{e}] \in \mathbb{R}^{d+p}$, $\tilde{\ell}(\boldsymbol{z}, \cdot, \boldsymbol{r}, \boldsymbol{e})$ is differentiable. Also $\tilde{\ell}(\boldsymbol{z}, \cdot, \cdot, \cdot)$ is continuous on $\mathbb{R}^{d+p} \times \mathcal{L}$. $\nabla_L \tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{r}, \boldsymbol{e}) = (L\boldsymbol{r} + \boldsymbol{e} - \boldsymbol{z})\boldsymbol{r}^\top$ is continuous on $\mathbb{R}^{d+p} \times \mathcal{L}$. $\forall L \in \mathcal{L}$, according to Assumption** $(A3)$, $\tilde{\ell}(\boldsymbol{z}, L, \cdot, \cdot)$ has a unique minimizer. Thus Lemma 2 applies and we prove that $\ell(\boldsymbol{z}, L)$ is differentiable in L and

$$\nabla_L \ell(\boldsymbol{z}, L) = (L\boldsymbol{r}^* + \boldsymbol{e}^* - \boldsymbol{z})\boldsymbol{r}^{*\top}.$$

Since every term in $\nabla_L \ell(\boldsymbol{z}, L)$ is uniformly bounded (Assumption $(A1)$ and Proposition 3), we conclude that the gradient of $\ell(\boldsymbol{z}, \cdot)$ is uniformly bounded, implying that $\ell(\boldsymbol{z}, L)$ is uniformly Lipschitz w.r.t. L.

Corollary 5 *Let $f_t(L)$ be the empirical loss function defined in (2.8). Then $f_t(L)$ is uniformly bounded and Lipschitz.*

Proof As $\ell(z, L)$ can be uniformly bounded (Corollary 1), we derive the uniform boundedness of $f_t(L)$. Let $U \in \frac{1}{2}\|L\|_{2,\infty}$. By computing the subgradient of $f_t(L)$ at L, we have

$$\begin{aligned}\left\|\nabla_L f_t(L)\right\|_F &= \left\|\frac{1}{t}\sum_{i=1}^{t}\nabla_L \ell(z_i, L) + \frac{\lambda_1}{t}U\right\|_F \\ &\le \frac{1}{t}\sum_{i=1}^{t}\left\|(L\boldsymbol{r}_i + \boldsymbol{e}_i - z_i)\boldsymbol{r}_i^\top\right\|_F + \frac{\lambda_1}{t}\|L\|_F \\ &= \frac{1}{t}\sum_{i=1}^{t}\left\|L\boldsymbol{r}_i\boldsymbol{r}_i^\top + (\boldsymbol{e}_i - z_i)\boldsymbol{r}_i^\top\right\|_F + \frac{\lambda_1}{t}\|L\|_F \\ &\le \frac{1}{t}\sum_{i=1}^{t}\left(\|L\|_F \cdot \left\|\boldsymbol{r}_i\boldsymbol{r}_i^\top\right\|_F + \left\|(\boldsymbol{e}_i - z_i)\boldsymbol{r}_i^\top\right\|_F\right) + \frac{\lambda_1}{t}\|L\|_F\,.\end{aligned}$$

Note that all the terms (i.e. z_i, L, $\boldsymbol{r}_i$, $\boldsymbol{e}_i$) in the right hand inequality are uniformly bounded. Thus, we say that the subgradient of $f_t(L)$ is uniformly bounded and $f_t(L)$ is uniformly Lipschitz.

Proposition 9 *Let $f_t(L)$ and $f(L)$ be the empirical and expected loss functions we defined in (2.8) and (4.1). Then we have*

$$\mathbb{E}\left[\sqrt{t}\,\|f_t - f\|_\infty\right] = O(1).$$

Proof Based on Propositions 3 and 4, we argue that the set of measurable functions $\{\ell(z, L), L \in \mathcal{L}\}$ is P-Donsker (defined in Lemma 1). From Corollary 1, we know that $\ell(z, L)$ can be uniformly bounded by a constant, say κ_c. Also note that from the definition of $\ell(z, L)$ (see (2.9)), it is always non-negative. Thus, we have

$$\ell^2(z, L) \le \kappa_c^2,$$

implying the uniform boundedness of $\mathbb{E}[\ell^2(z, L)]$. Thus, Lemma 1 applies and we have

$$\mathbb{E}\left[\sup_{\ell}|\sqrt{t}(f_t - f)|\right] = O(1).$$

We are ready to prove the convergence of $g_t(L_t)$, which requires to justify that the stochastic process $\{g_t(L_t)\}_{t=1}^{\infty}$ is a quasi-martingale, defined as follows:

Lemma 3 (Sufficient condition of convergence for a stochastic process *(Bottou 1998)*) *Let $(\Omega, \mathcal{F}, P)$ be a measurable probability space, u_t, for $t \ge 0$, be the realization of a stochastic process and $\mathcal{F}_t$ be the filtration by the past information at time t. Let*

$$\delta_t = \begin{cases} 1 & if\ \mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right] > 0, \\ 0 & otherwise. \end{cases}$$

If for all t, $u_t \ge 0$ and $\sum_{t=1}^{\infty}\mathbb{E}[\delta_t(u_{t+1} - u_t)] < \infty$, then u_t is a quasi-martingale and converges almost surely. Moreover,

$$\sum_{t=1}^{\infty}\left|\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right]\right| < +\infty\ a.s.$$

Theorem 4 (Convergence of the surrogate function $g_t(L_t)$) *The surrogate function $g_t(L_t)$ we defined in (3.1) converges almost surely, where L_t is the solution produced by Algorithm 1.*

Proof For convenience, let us first define the stochastic positive process

$$u_t = g_t(L_t) \geq 0.$$

We consider the difference between u_{t+1} and u_t:

$$\begin{aligned} u_{t+1} - u_t &= g_{t+1}(L_{t+1}) - g_t(L_t) \\ &= g_{t+1}(L_{t+1}) - g_{t+1}(L_t) + g_{t+1}(L_t) - g_t(L_t) \\ &= g_{t+1}(L_{t+1}) - g_{t+1}(L_t) + \frac{1}{t+1}\ell(z_{t+1}, L_t) - \frac{1}{t+1} g_t(L_t) \\ &= g_{t+1}(L_{t+1}) - g_{t+1}(L_t) + \frac{f_t(L_t) - g_t(L_t)}{t+1} + \frac{\ell(z_{t+1}, L_t) - f_t(L_t)}{t+1}. \end{aligned} \tag{8.3}$$

As L_{t+1} minimizes $g_{t+1}(L)$, we have

$$g_{t+1}(L_{t+1}) - g_{t+1}(L_t) \leq 0.$$

As $g_t(L_t)$ is the surrogate function of $f_t(L_t)$, we have

$$f_t(L_t) - g_t(L_t) \leq 0.$$

Thus,

$$u_{t+1} - u_t \leq \frac{\ell(z_{t+1}, L_t) - f_t(L_t)}{t+1}. \tag{8.4}$$

Let us consider the filtration of the past information $\mathcal{F}_t$ and take the expectation of (8.4) conditioned on $\mathcal{F}_t$:

$$\begin{aligned} \mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t] &\leq \frac{\mathbb{E}[\ell(z_{t+1}, L_t) \mid \mathcal{F}_t] - f_t(L_t)}{t+1} \\ &\leq \frac{f(L_t) - f_t(L_t)}{t+1} \\ &= \frac{f(L_t) - f_t'(L_t) - \frac{\lambda_1}{2t}\|L_t\|_{2,\infty}^2}{t+1} \\ &\leq \frac{\|f - f_t'\|_\infty}{t+1} - \frac{\lambda_1}{2t(t+1)}\|L_t\|_{2,\infty}^2 \\ &\leq \frac{\|f - f_t'\|_\infty}{t+1}, \end{aligned} \tag{8.5}$$

where

$$f_t'(L) = \frac{1}{t}\sum_{i=1}^{t} \ell(z_i, L).$$

Note that

$$f'(L) = \lim_{t\to\infty} f_t'(L) = \mathbb{E}_z[\ell(z, L)] = f(L).$$

From Proposition 5, we have

$$\mathbb{E}\Big[\Big\|\sqrt{t}(f_t' - f')\Big\|_\infty\Big] = O(1).$$

Also note that according to Proposition 3, we have $\|L_t\|_F \le L_{\max}$. Thus, considering the positive part of $\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]$ in (8.5) and taking the expectation, we have

$$\mathbb{E}\left[\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right]^+\right] = \mathbb{E}\left[\max\left\{\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right], 0\right\}\right] \le \frac{\kappa}{\sqrt{t}(t+1)},$$

where κ is a constant.

Therefore, defining the set $\mathcal{T} = \{t \mid \mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t] > 0\}$ and

$$\delta_t = \begin{cases} 1 & \text{if } t \in \mathcal{T}, \\ 0 & \text{otherwise,} \end{cases}$$

we have

$$\begin{aligned} \sum_{t=1}^{\infty} \mathbb{E}\left[\delta_t (u_{t+1} - u_t)\right] &= \sum_{t \in \mathcal{T}} \mathbb{E}\left[(u_{t+1} - u_t)\right] \\ &= \sum_{t \in \mathcal{T}} \mathbb{E}\left[\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right]\right] \\ &= \sum_{t=1}^{\infty} \mathbb{E}\left[\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right]^+\right] \\ &< +\infty. \end{aligned}$$

According to Lemma 3, we conclude that $g_t(L_t)$ is a quasi-martingale and converges almost surely. Moreover,

$$\sum_{t=1}^{\infty} \left|\mathbb{E}\left[u_{t+1} - u_t \mid \mathcal{F}_t\right]\right| < +\infty \; a.s. \tag{8.6}$$

8.4 Proof for Stage III

We now show that $g_t(L_t)$ and $f(L_t)$ converge to the same limit almost surely. Consequently, $f(L_t)$ converges almost surely. First, we prove that $b_t \overset{\text{def}}{=} g_t(L_t) - f_t(L_t)$ converges to 0 almost surely. We utilize the lemma from Mairal et al. (2010) for the proof.

Lemma 4 (Lemma 8 from *Mairal et al. (2010)*) *Let a_t, b_t be two real sequences such that for all t, $a_t \ge 0$, $b_t \ge 0$, $\sum_{t=1}^{\infty} a_t = \infty$, $\sum_{t=1}^{\infty} a_t b_t < \infty$, $\exists K > 0$, such that $|b_{t+1} - b_t| < K a_t$. Then, $\lim_{t \to +\infty} b_t = 0$.*

We notice that another sequence $\{a_t\}_{t=1}^{\infty}$ should be constructed in Lemma 4. Here, we take the $a_t = \frac{1}{t} \ge 0$, which satisfies the condition $\sum_{t=1}^{\infty} a_t = \infty$. Next, we need to show that $|b_{t+1} - b_t| < K a_t$, where K is a constant. To do this, we alternatively show that $|b_{t+1} - b_t|$ can be upper bounded by $\|L_{t+1} - L_t\|_F$, which can be further bounded by $K a_t$.

Proposition 10 *Let $\{L_t\}$ be the basis sequence produced by the Algorithm 1. Then,*

$$\|L_{t+1} - L_t\|_F = O\left(\frac{1}{t}\right).$$

Proof Let us define

$$\hat{g}_t(L) = \frac{1}{t}\left(\frac{1}{2}\mathrm{Tr}\left(L^\top L A_t\right) - \mathrm{Tr}\left(L^\top B_t\right)\right) + \frac{\lambda_1}{2t}\|L\|_{2,\infty}^2. \tag{8.7}$$

According the strong convexity of A_t (Assumption $(A2)$), and the convexity of $\|L\|_{2,\infty}^2$, we can derive the strong convexity of $\hat{g}_t(L)$. That is,

$$\hat{g}_t(L_{t+1}) - \hat{g}_t(L_t) \geq \langle U_t, L_{t+1} - L_t \rangle + \frac{\beta_1}{2} \|L_{t+1} - L_t\|_F^2 , \tag{8.8}$$

where $U_t \in \partial \hat{g}_t(L_t)$. As L_t is the minimizer of $\hat{g}_t$, we have

$$0 \in \partial \hat{g}_t(L_t).$$

Let U_t be the zero matrix. Then we have

$$\hat{g}_t(L_{t+1}) - \hat{g}_t(L_t) \geq \frac{\beta_1}{2} \|L_{t+1} - L_t\|_F^2 . \tag{8.9}$$

On the other hand,

$$\begin{aligned} \hat{g}_t(L_{t+1}) - \hat{g}_t(L_t) &= \hat{g}_t(L_{t+1}) - \hat{g}_{t+1}(L_{t+1}) + \hat{g}_{t+1}(L_{t+1}) \\ &\quad - \hat{g}_{t+1}(L_t) + \hat{g}_{t+1}(L_t) - \hat{g}_t(L_t) \\ &\leq \hat{g}_t(L_{t+1}) - \hat{g}_{t+1}(L_{t+1}) + \hat{g}_{t+1}(L_t) - \hat{g}_t(L_t). \end{aligned} \tag{8.10}$$

Note that the inequality is derived by the fact that $\hat{g}_{t+1}(L_{t+1}) - \hat{g}_{t+1}(L_t) \leq 0$, as L_{t+1} is the minimizer of $\hat{g}_{t+1}(L)$. Let us denote $\hat{g}_t(L) - \hat{g}_{t+1}(L)$ by $G_t(L)$. We have

$$\begin{aligned} G_t(L) = \frac{1}{t}\left(\frac{1}{2}\mathrm{Tr}\left(L^\top L A_t\right) - \mathrm{Tr}\left(L^\top B_t\right)\right) - \frac{1}{t+1}\left(\frac{1}{2}\mathrm{Tr}\left(L^\top L A_{t+1}\right) - \mathrm{Tr}\left(L^\top B_{t+1}\right)\right) \\ + \frac{\lambda_1}{2t}\|L\|_{2,\infty}^2 - \frac{\lambda_1}{2(t+1)}\|L\|_{2,\infty}^2 . \end{aligned}$$

By a simple calculation, we have the gradient of $G_t(L)$:

$$\begin{aligned} \nabla G_t(L) &= \frac{1}{t}(L A_t - B_t) - \frac{1}{t+1}(L A_{t+1} - B_{t+1}) + \left(\frac{1}{t} - \frac{1}{t+1}\right)\lambda_1 U \\ &= \frac{1}{t}\left(L(A_t - \frac{t}{t+1}A_{t+1}) + \frac{t}{t+1}B_{t+1} - B_t + \frac{\lambda_1}{t+1}U\right), \end{aligned}$$

where $U \in \partial \|L\|_{2,\infty}^2$. We then compute the Frobenius norm of the gradient of $G_t(L)$:

$$\begin{aligned} \|\nabla G_t(L)\|_F &\leq \frac{1}{t}\left(\left\|L(A_t - \frac{t}{t+1}A_{t+1})\right\|_F + \left\|\frac{t}{t+1}B_{t+1} - B_t\right\|_F + \frac{\lambda_1}{t+1}\|L\|_F\right) \\ &\leq \frac{1}{t}\left(\|L\|_F \cdot \left\|A_t - \frac{t}{t+1}A_{t+1}\right\|_F + \left\|\frac{t}{t+1}B_{t+1} - B_t\right\|_F + \frac{\lambda_1}{t+1}\|L\|_F\right) \\ &= \frac{1}{t}\Big(\|L\|_F \cdot \left\|\frac{1}{t+1}A_t - \frac{t}{t+1}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}^\top\right\|_F \\ &\quad + \left\|\frac{1}{t+1}B_t - \frac{t}{t+1}(\boldsymbol{z}_{t+1} - \boldsymbol{e}_{t+1})\boldsymbol{r}_{t+1}^\top\right\|_F + \frac{\lambda_1}{t+1}\|L\|_F\Big). \end{aligned} \tag{8.11}$$

According to the first order Taylor expansion,

$$\begin{aligned} G_t(L_{t+1}) - G_t(L_t) &= \mathrm{Tr}\left((L_{t+1} - L_t)^\top \nabla G_t(\alpha L_t + (1-\alpha)L_{t+1})\right) \\ &\leq \|L_{t+1} - L_t\|_F \cdot \|\nabla G_t(\alpha L_t + (1-\alpha)L_{t+1})\|_F , \end{aligned}$$

where α is a constant between 0 and 1. According to Proposition 3, L_t and L_{t+1} are uniformly bounded, so $\alpha L_t + (1-\alpha)L_{t+1}$ is uniformly bounded. According to Proposition 3, $\frac{1}{t+1}A_t$,

$\frac{t}{t+1}\boldsymbol{r}_{t+1}\boldsymbol{r}_{t+1}^\top$, $\frac{1}{t+1}B_t$ and $\frac{t}{t+1}\left(\boldsymbol{z}_{t+1} - \boldsymbol{e}_{t+1}\right)\boldsymbol{r}_{t+1}^\top$ are all uniformly bounded. Thus, there exists a constant c, such that

$$\left\|\nabla G_t\left(\alpha L_t + (1-\alpha) L_{t+1}\right)\right\|_F \le \frac{c}{t},$$

resulting that

$$G_t(L_{t+1}) - G_t(L_t) \le \frac{c}{t}\left\|L_{t+1} - L_t\right\|_F .$$

Applying this property in (8.10), we have

$$\hat{g}_t(L_{t+1}) - \hat{g}_t(L_t) \le G_t(L_{t+1}) - G_t(L_t) \le \frac{c}{t}\left\|L_{t+1} - L_t\right\|_F . \tag{8.12}$$

From (8.9) and (8.12), we conclude that

$$\left\|L_{t+1} - L_t\right\|_F \le \frac{2c}{\beta_1}\cdot\frac{1}{t}. \tag{8.13}$$

Theorem 5 (Convergence of the empirical and expected loss) *Let $\{f(L_t)\}_{t=1}^{\infty}$ be the sequence of the expected loss where $\{L_t\}_{t=1}^{\infty}$ be the sequence of the solutions produced by the Algorithm 1. Also for any $t > 0$, denote $g_t(L_t) - f_t(L_t)$ by b_t. Then,*

1. *The sequence $\{b_t\}_{t=1}^{\infty}$ converges almost surely to 0.*
2. *The sequence of the empirical loss $\{f_t(L_t)\}_{t=1}^{\infty}$ converges almost surely.*
3. *The sequence of the expected loss $\{f(L_t)\}_{t=1}^{\infty}$ converges almost surely to the same limit of the surrogate $\{g_t(L_t)\}_{t=1}^{\infty}$.*

Proof We start our proof by deriving an upper bound for $g_t(L_t) - f_t(L_t)$.

Step 1 According to (8.3),

$$\begin{aligned}\frac{b_t}{t+1} &= g_{t+1}(L_{t+1}) - g_{t+1}(L_t) + \frac{\ell(z_{t+1}, L_t) - f_t(L_t)}{t+1} + u_t - u_{t+1}\\ &\le \frac{\ell(z_{t+1}, L_t) - f_t(L_t)}{t+1} + u_t - u_{t+1}.\end{aligned}$$

Taking the expectation conditioned on the past information $\mathcal{F}_t$ in the above equation, and note that

$$\begin{aligned}\mathbb{E}\left[\frac{b_t}{t+1}\;\Big|\;\mathcal{F}_t\right] &= \frac{g_t(L_t) - f_t(L_t)}{t+1},\\ \mathbb{E}\left[\frac{\ell(z_{t+1}, L_t) - f_t(L_t)}{t+1}\;\Big|\;\mathcal{F}_t\right] &= \frac{f(L_t) - f_t(L_t)}{t+1},\end{aligned}$$

we have

$$\frac{b_t}{t+1} \le \frac{f(L_t) - f_t(L_t)}{t+1} + \mathbb{E}\left[u_t - u_{t+1} \mid \mathcal{F}_t\right].$$

Thus,

$$\sum_{t=1}^{\infty}\frac{b_t}{t+1} \le \sum_{t=1}^{\infty}\frac{f(L_t) - f_t(L_t)}{t+1} + \sum_{t=1}^{\infty}\mathbb{E}\left[u_t - u_{t+1} \mid \mathcal{F}_t\right].$$

According to the central limit theorem, $\sqrt{t}(f(L_t) - f_t(L_t))$ is bounded almost surely. Also, from (8.6),

$$\sum_{t=1}^{\infty} \mathbb{E}\left[u_t - u_{t+1} \mid \mathcal{F}_t\right] \leq \sum_{t=1}^{\infty} \left|\mathbb{E}\left[u_t - u_{t+1} \mid \mathcal{F}_t\right]\right| < +\infty.$$

Thus,

$$\sum_{t=1}^{\infty} \frac{b_t}{t+1} < +\infty.$$

Step 2 We examine the difference between b_{t+1} and b_t:

$$\begin{aligned}
|b_{t+1} - b_t| &= |g_{t+1}(L_{t+1}) - f_{t+1}(L_{t+1}) - g_t(L_t) + f_t(L_t)| \\
&\leq |g_{t+1}(L_{t+1}) - g_t(L_t)| + |f_{t+1}(L_{t+1}) - f_t(L_t)| \\
&= |g_{t+1}(L_{t+1}) - g_t(L_{t+1}) + g_t(L_{t+1}) - g_t(L_t)| \\
&\quad + |f_{t+1}(L_{t+1}) - f_t(L_{t+1}) + f_t(L_{t+1}) - f_t(L_t)| \\
&\leq |g_{t+1}(L_{t+1}) - g_t(L_{t+1})| + |g_t(L_{t+1}) - g_t(L_t)| \\
&\quad + |f_{t+1}(L_{t+1}) - f_t(L_{t+1})| + |f_t(L_{t+1}) - f_t(L_t)| \\
&= \left|\frac{1}{t+1}\ell(z_{t+1}, L_{t+1}) - \frac{1}{t+1}g_t(L_{t+1})\right| + |g_t(L_{t+1}) - g_t(L_t)| \\
&\quad + \left|\frac{1}{t+1}\ell(z_{t+1}, L_{t+1}) - \frac{1}{t+1}f_t(L_{t+1})\right| + |f_t(L_{t+1}) - f_t(L_t)|.
\end{aligned}$$

According to Corollaries 1 and 2, we know that there exist constant κ_1 and κ_2 that are uniformly over t, such that

$$\begin{aligned}
|g_t(L_{t+1}) - g_t(L_t)| &\leq \kappa_1 \|L_{t+1} - L_t\|_F, \\
|f_t(L_{t+1}) - f_t(L_t)| &\leq \kappa_2 \|L_{t+1} - L_t\|_F.
\end{aligned}$$

Combing with Proposition 6, there exists a constant κ_3 that is uniformly over t, such that

$$|g_t(L_{t+1}) - g_t(L_t)| + |f_t(L_{t+1}) - f_t(L_t)| \leq \frac{\kappa_3}{t}.$$

As we shown, $\ell(z_{t+1}, L_{t+1})$, $g_t(L_{t+1})$ and $f_t(L_{t+1})$ are all uniformly bounded. Therefore, there exists a constant κ_4, such that

$$|\ell(z_{t+1}, L_{t+1}) - g_t(L_{t+1})| + |\ell(z_{t+1}, L_{t+1}) - f_t(L_t + 1)| \leq \kappa_4.$$

Finally, we have

$$b_{t+1} - b_t \leq \frac{\kappa_4}{t+1} + \frac{\kappa_3}{t} \leq \frac{\kappa_5}{t},$$

where κ_5 is a constant that is uniformly over t.

Applying Lemma 4, we conclude that $\{b_t\}$ converges to zero. That is,

$$\lim_{t \to +\infty} g_t(L_t) - f_t(L_t) = 0. \tag{8.14}$$

In Theorem 2, we have shown that $g_t(L_t)$ converges almost surely. This implies that $f_t(L_t)$ also converges almost surely to the same limit of $g_t(L_t)$.

According to the central limit theorem, $\sqrt{t}(f(L_t) - f_t(L_t)$ is bounded, implying

$$\lim_{t \to +\infty} f(L_t) - f_t(L_t) = 0, \quad a.s.$$

Thus, we conclude that $f(L_t)$ converges almost surely to the same limit of $f_t(L_t)$ (or, $g_t(L_t)$).

8.5 Finalizing the Proof

According to Theorem 3, we can see that $g_t(L_t)$ and $f(L_t)$ converge to the same limit almost surely. Let t tends to infinity, as L_t is uniformly bounded (Proposition 3), the term $\frac{\lambda_1}{2t} \|L_t\|_{2,\infty}^2$ in $g_t(L_t)$ vanishes. Thus $g_t(L_t)$ becomes differentiable. On the other hand, we have the following proposition about the gradient of $f(L)$.

Proposition 11 (Subgradient of $f(L)$) *Let $f(L)$ be the expected loss function defined in* (4.1)*. Then, $f(L)$ is continuously differentiable and $\nabla f(L) = \mathbb{E}_z[\nabla_L \ell(z, L)]$. Moreover, $\nabla f(L)$ is uniformly Lipschitz on $\mathcal{L}$.*

Proof Since $\ell(z, L)$ is continuously differentiable (Proposition 4), $f(L)$ is continuously differentiable and $\nabla f(L) = \mathbb{E}_z[\nabla_L \ell(z, L)]$.

Now we prove the second claim. Let us consider a matrix L and a sample z, and denote $\boldsymbol{r}^*(z, L)$ and $\boldsymbol{e}^*(z, L)$ as the optimal solutions for (2.9).

Step 1 First, $\tilde{\ell}(z, L, \boldsymbol{r}, \boldsymbol{e})$ is continuous in z, L, $\boldsymbol{r}$ and $\boldsymbol{e}$, and has a unique minimizer. This implies that $\boldsymbol{r}^*(z, L)$ and $\boldsymbol{e}^*(z, L)$ is continuous in z and L.

Let us denote Λ as the set of the indices such that $\forall j \in \Lambda$, $\boldsymbol{e}_j^* \neq 0$. According to the first order optimal condition for (3.2) w.r.t $\boldsymbol{e}$, we have

$$z - L\boldsymbol{r} - \boldsymbol{e} \in \lambda_2 \partial \|\boldsymbol{e}\|_1 ,$$

implying

$$\left|(z - L\boldsymbol{r} - \boldsymbol{e})_j\right| = \lambda_2, \ \forall j \in \Lambda.$$

Since $z - L\boldsymbol{r} - \boldsymbol{e}$ is continuous in z and L, we consider a small perturbation of (z, L) in one of their open neighborhood V, such that for all $(z', L') \in V$, we have if $j \notin \Lambda$, then $\left|(z' - L'\boldsymbol{r}^{*\prime} - \boldsymbol{e}^{*\prime})_j\right| < \lambda_2$ and $\boldsymbol{e}^{*\prime}_j = 0$, where $\boldsymbol{r}^{*\prime} = \boldsymbol{r}^*(z', L')$ and $\boldsymbol{e}^{*\prime} = \boldsymbol{e}^*(z', L')$. That is, the support set of $\boldsymbol{e}^*$ does not change.

Let us denote $D = [L \ I]$ and $\boldsymbol{b} = [\boldsymbol{r}; \ \boldsymbol{e}]$ and consider the function

$$\tilde{\ell}(z, L_\Lambda, \boldsymbol{b}_\Lambda) \overset{\text{def}}{=} \frac{1}{2} \left\|z - D_\Lambda \boldsymbol{b}_\Lambda\right\|_2^2 + \lambda_2 \left\|[0 \ I]\boldsymbol{b}_\Lambda\right\|_1 .$$

According to Assumption (8.3), $\tilde{\ell}(z, L_\Lambda, \cdot)$ is strongly convex with a Hessian lower-bounded by a positive constant κ_1. Thus,

$$\tilde{\ell}(z, L_\Lambda, \boldsymbol{b}'^*_\Lambda) - \tilde{\ell}(z, L_\Lambda, \boldsymbol{b}^*_\Lambda) \geq \kappa_1 \left\|\boldsymbol{b}_\Lambda - \boldsymbol{b}'_\Lambda\right\|_2^2 = \kappa_1 \left(\left\|\boldsymbol{r}^* - \boldsymbol{r}'^*\right\|_2^2 + \left\|\boldsymbol{e}^*_\Lambda - \boldsymbol{e}'^*_\Lambda\right\|_2^2\right). \tag{8.15}$$

Step 2 We shall prove that $\tilde{\ell}(z, L, \cdot) - \tilde{\ell}(z', L', \cdot)$ is Lipschitz w.r.t. $\boldsymbol{b}$.

$$\begin{aligned} &2\left(\tilde{\ell}(z, L, \boldsymbol{b}) - \tilde{\ell}(z', L', \boldsymbol{b})\right) - 2\left(\tilde{\ell}(z, L, \boldsymbol{b}') - \tilde{\ell}(z', L', \boldsymbol{b}')\right) \\ &= \|z - D\boldsymbol{b}\|_2^2 - \|z - D\boldsymbol{b}'\|_2^2 + \|z' - D'\boldsymbol{b}'\|_2^2 - \|z' - D'\boldsymbol{b}\|_2^2 \end{aligned}$$

$$
\begin{aligned}
&= 2z^\top D(\boldsymbol{b}' - \boldsymbol{b}) + \boldsymbol{b}^\top D^\top D\boldsymbol{b} - \boldsymbol{b}'^\top D^\top D\boldsymbol{b}' - 2z'^\top D'(\boldsymbol{b}' - \boldsymbol{b})\\
&\quad - \boldsymbol{b}^\top D'^\top D'\boldsymbol{b} + \boldsymbol{b}'^\top D'^\top D'\boldsymbol{b}'\\
&= 2\left[(z^\top D - z'^\top D')(\boldsymbol{b}' - \boldsymbol{b})\right]\\
&\quad + \left[\boldsymbol{b}^\top D^\top D\boldsymbol{b} - \boldsymbol{b}^\top D'^\top D'\boldsymbol{b} + \boldsymbol{b}'^\top D'^\top D'\boldsymbol{b}' - \boldsymbol{b}'^\top D^\top D\boldsymbol{b}'\right].
\end{aligned}
$$

For the first term,

$$
\begin{aligned}
(z^\top D - z'^\top D')(\boldsymbol{b}' - \boldsymbol{b}) &= (z^\top D - z^\top D' + z^\top D' - z'^\top D'^\top)(\boldsymbol{b}' - \boldsymbol{b})\\
&= \left(z^\top (D - D') + (z^\top - z'^\top)D'\right)(\boldsymbol{b}' - \boldsymbol{b}).
\end{aligned}
$$

As each sample is bounded, D is bounded (as L is bounded), so the ℓ_2 norm of the first term can be bounded as follows:

$$
\begin{aligned}
&\left\|(z^\top D - z'^\top D')(\boldsymbol{b}' - \boldsymbol{b})\right\|_2\\
&\quad = \left\|\left(z^\top (D - D') + (z^\top - z'^\top)D'\right)(\boldsymbol{b}' - \boldsymbol{b})\right\|_2\\
&\quad \le \left(\|z\|_2\|D - D'\|_F + \|z - z'\|_2\|D'\|_F\right) \cdot \left\|\boldsymbol{b}' - \boldsymbol{b}\right\|_2\\
&\quad \le \left(c_1 \left\|D - D'\right\|_F + c_2\|z - z'\|_2\right) \cdot \left\|\boldsymbol{b}' - \boldsymbol{b}\right\|_2 .
\end{aligned}
\tag{8.16}
$$

For the second term, we have

$$
\begin{aligned}
&\boldsymbol{b}^\top D^\top D\boldsymbol{b} - \boldsymbol{b}^\top D'^\top D'\boldsymbol{b} + \boldsymbol{b}'^\top D'^\top D'\boldsymbol{b}' - \boldsymbol{b}'^\top D^\top D\boldsymbol{b}'\\
&\quad = \boldsymbol{b}^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b} - \boldsymbol{b}'^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b}'\\
&\quad = \boldsymbol{b}^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b} - \boldsymbol{b}^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b}' + \boldsymbol{b}^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b}'\\
&\quad - \boldsymbol{b}'^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b}'\\
&\quad = \boldsymbol{b}^\top \left(D^\top D - D'^\top D'\right)\left(\boldsymbol{b} - \boldsymbol{b}'\right) + \left(\boldsymbol{b} - \boldsymbol{b}'\right)^\top \left(D^\top D - D'^\top D'\right)\boldsymbol{b}'\\
&\quad = \boldsymbol{b}^\top \left(D^\top D - D^\top D' + D^\top D' - D'^\top D'\right)\left(\boldsymbol{b} - \boldsymbol{b}'\right)\\
&\qquad + \left(\boldsymbol{b} - \boldsymbol{b}'\right)^\top \left(D^\top D - D^\top D' + D^\top D' - D'^\top D'\right)\boldsymbol{b}'\\
&\quad = \boldsymbol{b}^\top \left(D^\top \left(D - D'\right) + \left(D^\top - D'\right)D'\right)\left(\boldsymbol{b} - \boldsymbol{b}'\right)\\
&\qquad + \left(\boldsymbol{b} - \boldsymbol{b}'\right)^\top \left(D^\top \left(D - D'\right) + \left(D^\top - D'\right)D'\right)\boldsymbol{b}'.
\end{aligned}
$$

Since D is bounded and $\boldsymbol{b}$ is bounded, the second term can be bounded as follows:

$$
\begin{aligned}
&\left\|\boldsymbol{b}^\top D^\top D\boldsymbol{b} - \boldsymbol{b}^\top D'^\top D'\boldsymbol{b} + \boldsymbol{b}'^\top D'^\top D'\boldsymbol{b}' - \boldsymbol{b}'^\top D^\top D\boldsymbol{b}'\right\|_2\\
&\quad = \Big\|\boldsymbol{b}^\top \left(D^\top \left(D - D'\right) + \left(D^\top - D'^\top\right)D'\right)\left(\boldsymbol{b} - \boldsymbol{b}'\right)\\
&\qquad + \left(\boldsymbol{b} - \boldsymbol{b}'\right)^\top \left(D^\top \left(D - D'\right) + \left(D^\top - D'^\top\right)D'\right)\boldsymbol{b}'\Big\|\\
&\quad \le c_3 \left\|D - D'\right\|_F \cdot \left\|\boldsymbol{b} - \boldsymbol{b}'\right\|_2 .
\end{aligned}
\tag{8.17}
$$

Combining (8.16) and (8.17), we prove that the function $\tilde{\ell}(\boldsymbol{z}, L, \cdot) - \tilde{\ell}(\boldsymbol{z}', L', \cdot)$ is Lipschitz:

$$\begin{aligned}
&\left(\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{b}) - \tilde{\ell}(\boldsymbol{z}', L', \boldsymbol{b})\right) - \left(\tilde{\ell}(\boldsymbol{z}, L, \boldsymbol{b}') - \tilde{\ell}(\boldsymbol{z}', L', \boldsymbol{b}')\right) \\
&\quad \le \left((c_1 + c_3)\left\|D - D'\right\|_F + c_2\left\|\boldsymbol{z} - \boldsymbol{z}'\right\|_2\right)\left\|\boldsymbol{b} - \boldsymbol{b}'\right\|_2 \\
&\quad = \left((c_1 + c_3)\left\|D - D'\right\|_F + c_2\left\|\boldsymbol{z} - \boldsymbol{z}'\right\|_2\right)\sqrt{\|\boldsymbol{r} - \boldsymbol{r}'\|_2^2 + \|\boldsymbol{e} - \boldsymbol{e}'\|_2^2}. \qquad (8.18)
\end{aligned}$$

Step 3 According to (8.15) and (8.18), and the fact that $\boldsymbol{b}'^*$ minimizes $\tilde{\ell}(\boldsymbol{z}', L', \cdot)$, we have

$$\begin{aligned}
&\kappa_1\left(\left\|\boldsymbol{r}^* - \boldsymbol{r}'^*\right\|_2^2 + \left\|\boldsymbol{e}_\Lambda^* - \boldsymbol{e}_\Lambda'^*\right\|_2^2\right) \\
&\quad \le \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda'^*) - \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda^*) \\
&\quad = \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda'^*) - \tilde{\ell}(\boldsymbol{z}', L_\Lambda', \boldsymbol{b}_\Lambda^*) + \tilde{\ell}(\boldsymbol{z}', L_\Lambda', \boldsymbol{b}_\Lambda^*) - \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda^*) \\
&\quad \le \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda'^*) - \tilde{\ell}(\boldsymbol{z}', L_\Lambda', \boldsymbol{b}_\Lambda'^*) + \tilde{\ell}(\boldsymbol{z}', L_\Lambda', \boldsymbol{b}_\Lambda^*) - \tilde{\ell}(\boldsymbol{z}, L_\Lambda, \boldsymbol{b}_\Lambda^*) \\
&\quad \le \left((c_1 + c_3)\left\|D - D'\right\|_F + c_2\left\|\boldsymbol{z} - \boldsymbol{z}'\right\|_2\right)\sqrt{\|\boldsymbol{r}^* - \boldsymbol{r}'^*\|_2^2 + \left\|\boldsymbol{e}_\Lambda^* - \boldsymbol{e}_\Lambda'^*\right\|_2^2}.
\end{aligned}$$

Therefore, $\boldsymbol{r}^*(\boldsymbol{z}, L)$ and $\boldsymbol{e}^*(\boldsymbol{z}, L)$ are Lipschitz, which concludes the proof.

Finally, taking a first order Taylor expansion for $f(L_t)$ and $g_t(L_t)$, we can show that the gradient of $f(L_t)$ equals to that of $g_t(L_t)$ when t tends to infinity. Since L_t is the minimizer for $g_t(L)$, we know that the gradient of $f(L_t)$ vanishes. Therefore, we have proved Theorem 1.

Proof According to Proposition 3, the sequences $\{\frac{1}{t}A_t\}$ and $\{\frac{1}{t}B_t\}$ are uniformly bounded. Then, there exist sub-sequences of $\{\frac{1}{t}A_t\}$ and $\{\frac{1}{t}B_t\}$ that converge to A_∞ and B_∞ respectively. In that case, L_t converges to L_∞. Let V be an arbitrary matrix in $\mathbb{R}^{p\times d}$, and $\{h_k\}$ be a positive sequence that converges to zero.

Since g_t is the surrogate function of f_t, for all t and k, we have

$$g_t(L_t + h_k V) \ge f_t(L_t + h_k V).$$

Let t tend to infinity:

$$g_\infty(L_\infty + h_k V) \ge f(L_\infty + h_k V).$$

Since L_t is uniformly bounded, when t tends to infinity, the term $\frac{\lambda_1}{2t}\left\|L_t\right\|_\infty^2$ will vanish. In this way, $g_t(\cdot)$ becomes differentiable. Also, the Lipschitz of $\nabla f(L)$ (proved in Proposition 11) implies that the second derivative of $f(L_t)$ can be uniformly bounded. And by a simple calculation, this also holds for $g_t(L_t)$. Thus, we can take the first order Taylor expansion even when t tends to infinity. Using a first order Taylor expansion, and note the fact that $g_\infty(L_\infty) = f(L_\infty)$, we have

$$\operatorname{Tr}\left(h_k V^\top \nabla g_\infty(L_\infty)\right) + o(h_k V) \ge \operatorname{Tr}\left(h_k V^\top \nabla f(L_\infty)\right) + o(h_k V).$$

Since $\{h_k\}$ is a positive sequence, by multiplying $\frac{1}{h_k\|V\|_F}$ on both side, it follows that

$$\operatorname{Tr}\left(\frac{1}{\|V\|_F}V^\top \nabla g_\infty(L_\infty)\right) + \frac{o(h_k V)}{h_k\|V\|_F} \ge \operatorname{Tr}\left(\frac{1}{\|V\|_F}V^\top \nabla f(L_\infty)\right) + \frac{o(h_k V)}{h_k\|V\|_F}.$$

Now let k tend to infinity:

$$\operatorname{Tr}\left(\frac{1}{\|V\|_F}V^\top \nabla g_\infty(L_\infty)\right) \ge \operatorname{Tr}\left(\frac{1}{\|V\|_F}V^\top \nabla f(L_\infty)\right).$$

Since the inequality holds for all matrix $V \in \mathbb{R}^{p \times d}$, it can easily show that

$$\nabla g_{\infty}(L_{\infty}) = \nabla f(L_{\infty}).$$

Since L_t always minimizes $g_t(\cdot)$, we have

$$\nabla f(L_{\infty}) = \nabla g_{\infty}(L_{\infty}) = 0,$$

which implies that when t tend to infinity, L_t is a stationary point of $f(\cdot)$.

References

Artač, M., Jogan, M., & Leonardis, A. (2002). Incremental PCA for on-line visual learning and recognition. In *Proceedings of the 16th international conference on pattern recognition* (Vol. 3, pp. 781–784).

Bertsekas, D. P. (1999). *Nonlinear programming*. Massachusetts: Athena Scientific.

Bhojanapalli, S., Kyrillidis, A., & Sanghavi, S. (2016). Dropping convexity for faster semi-definite optimization. In *Proceedings of the 29th conference on learning theory* (pp. 530–582).

Bonnans, J. F., & Shapiro, A. (1998). Optimization problems with perturbations: A guided tour. *SIAM Review*, *40*(2), 228–264.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line Learning in Neural Networks*, *17*(9), 142.

Burer, S., & Monteiro, R. D. C. (2005). Local minima and convergence in low-rank semidefinite programming. *Mathematical Programming*, *103*(3), 427–444.

Cai, J., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, *20*(4), 1956–1982.

Cai, T. T., & Zhou, W. (2013). A max-norm constrained minimization approach to 1-bit matrix completion. *Journal of Machine Learning Research*, *14*(1), 3619–3647.

Cai, T. T., & Zhou, W. X. (2016). Matrix completion via max-norm constrained optimization. *Electronic Journal of Statistics*, *10*(1), 1493–1525.

Candès, E. J., Li, X., Ma, Y., & Wright, J. (2011). Robust principal component analysis? *Journal of the ACM*, *58*(3), 11:1–11:37.

Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, *9*(6), 717–772.

Davenport, M. A., Plan, Y., van den Berg, E., & Wootters, M. (2014). 1-Bit matrix completion. *Information and Inference*, *3*(3), 189–223.

Donoho, D. L. (1995). De-noising by soft-thresholding. *IEEE Transactions on Information Theory*, *41*(3), 613–627.

Fazel, M., Hindi, H., & Boyd, S. P. (2001). A rank minimization heuristic with application to minimum order system approximation. In *Proceedings of the American control conference* (Vol. 6, pp. 4734–4739).

Feng, J., Xu, H., & Yan, S. (2013). Online robust PCA via stochastic optimization. In *Proceedings of the 27th annual conference on neural information processing systems* (pp. 404–412).

Foygel, R., Srebro, N., & Salakhutdinov, R. (2012). Matrix reconstruction with the local max norm. In *Proceedings of the 26th annual conference on neural information processing systems* (pp. 944–952).

Jalali, A., & Srebro, N. (2012). Clustering using max-norm constrained optimization. In *Proceedings of the 29th international conference on machine learning*.

Jolliffe, I. (2005). *Principal component analysis*. Hoboken: Wiley Online Library.

Klopp, O. (2014). Noisy low-rank matrix completion with general sampling distribution. *Bernoulli*, *20*(1), 282–303.

Lee, J. D., Recht, B., Salakhutdinov, R., Srebro, N., & Tropp, J. A. (2010). Practical large-scale optimization for max-norm regularization. In *Proceedings of the 24th annual conference on neural information processing systems* (pp. 1297–1305).

Liu, G., Lin, Z., & Yu, Y. (2010). Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th international conference on machine learning* (pp. 663–670).

Mairal, J. (2013). Stochastic majorization-minimization algorithms for large-scale optimization. In *Proceedings of the 27th annual conference on neural information processing systems* (pp. 2283–2291).

Mairal, J., Bach, F. R., Ponce, J., & Sapiro, G. (2010). Online learning for matrix factorization and sparse coding. *Journal of Machine Learning Research*, *11*, 19–60.

Neyshabur, B., Makarychev, Y., & Srebro, N. (2014). Clustering, hamming embedding, generalized LSH and the max norm. In *Proceedings of the 25th international conference on algorithmic learning theory* (pp. 306–320).

Orabona, F., Argyriou, A., & Srebro, N. (2012). PRISMA: PRoximal Iterative SMoothing Algorithm. CoRR abs/1206.2372.

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

Rennie, J. D. M., & Srebro, N. (2005). Fast maximum margin matrix factorization for collaborative prediction. In *Proceedings of the 22nd international conference on machine learning* (pp. 713–719).

Salakhutdinov, R., & Srebro, N. (2010). Collaborative filtering in a non-uniform world: Learning with the weighted trace norm. In *Proceedings of the 24th annual conference on neural information processing systems* (pp. 2056–2064).

Shen, J., Xu, H., & Li, P. (2014). Online optimization for max-norm regularization. In *Proceedings of the 28th annual conference on neural information processing systems* (pp. 1718–1726).

Srebro, N., Rennie, J. D. M., & Jaakkola, T. S. (2004). Maximum-margin matrix factorization. In *Proceedings of the 18th annual conference on neural information processing systems* (pp. 1329–1336).

Srebro, N., & Shraibman, A. (2005). Rank, trace-norm and max-norm. In *Proceedings of the 18th annual conference on learning theory* (pp. 545–560).

Van der Vaart, A. W. (2000). *Asymptotic statistics* (Vol. 3). Cambridge: Cambridge University Press.

Wang, H., & Banerjee, A. (2014). Randomized block coordinate descent for online and stochastic optimization. CoRR abs/1407.0107.

Xu, H., Caramanis, C., & Mannor, S. (2010). Principal component analysis with contaminated data: The high dimensional case. In *Proceedings of the 23rd conference on learning theory* (pp. 490–502).

Xu, H., Caramanis, C., & Mannor, S. (2013). Outlier-robust PCA: The high-dimensional case. *IEEE Transactions on Information Theory*, *59*(1), 546–572.

Xu, H., Caramanis, C., & Sanghavi, S. (2012). Robust PCA via outlier pursuit. *IEEE Transactions on Information Theory*, *58*(5), 3047–3064.

Zhou, Z., Li, X., Wright, J., Candès, E. J., & Ma, Y. (2010). Stable principal component pursuit. In *Proceedings of the 2010 IEEE international symposium on information theory* (pp. 1518–1522).