Abstract

Low-Rank Representation (LRR) is a popular tool to identify the data that are generated by a union of subspaces. However, the size of the regularized matrix of LRR is proportional to $n^2$, where $n$ is the number of samples, which hinders LRR for large scale problems. In this paper, we study how to scale up the LRR method accurately and memory efficiently. In particular, we propose an online implementation that reduces the memory cost from $O(n^2)$ to $O(pd)$, with $p$ being the ambient dimension and $d$ being some expected rank. There are two key techniques in our algorithm: one is to reformulate the nuclear norm to an equivalent matrix factorization form and the other is to introduce an auxiliary variable working as a basis dictionary of the underlying data, which combined together makes the problem amenable to stochastic optimization. We establish a theoretical guarantee that the sequence of solutions produced by our algorithm will converge to a stationary point asymptotically. Numerical experiments on subspace recovery and subspace clustering tasks demonstrate the efficacy and robustness of the proposed algorithm.

1 Introduction

The nuclear norm has been extensively studied in the last decade and finds successful applications in matrix completion, matrix decomposition, subspace clustering, robust principal component analysis (RPCA), to name a few.

However, in the face of big data, there are two challenges for solving a nuclear norm regularized problem which obstructs its practical success. First, it is usually computationally expensive to optimize such problems. For example, an SDP solver can obtain accurate enough solution but is not scalable to large matrices. The singular value thresholding technique needs to compute the SVD in each iteration, which is also not practical. Second, the memory cost of popular nuclear solvers is often proportional to the size of the regularized matrix. Although for specific problems such as matrix completion, it can be reduced to the number of observed entries, in general, the huge memory cost still hinders most of batch based nuclear norm solvers.

In this paper, we aim at scaling up the nuclear norm accurately and memory efficiently. Particularly, we address the worst case in terms of memory cost – the size of the nuclear norm regularized matrix is $n$ by $n$. Such a problem mainly emerges in low-rank subspace clustering and extensions of matrix completion problem. We study the following formulation (termed LRR) proposed by for the low-rank subspace clustering problem:

\[
\min_{X,E} \|X\|_* + \lambda \|E\|_1, \text{ s.t. } Z = ZX + E.
\]
Here, $Z \in \mathbb{R}^{p \times n}$ is the observation matrix with $n$ samples lying in a $p$ dimensional subspace and $E \in \mathbb{R}^{p \times n}$ is some sparse corruption. The program pursues a low-rank representation $X \in \mathbb{R}^{n \times n}$ among all candidates, each of which can be represented as a linear combination of the atoms in the dictionary $Z$. Note that in the above formulation, there is typically no sparsity assumption on $X$. Hence, most recently developed nuclear norm solvers [21, 19, 17] are not applicable to large scale LRR as the memory cost is $O(n^2)$; see Section 1.1 for a detailed discussion on this issue. Our algorithm, in contrast, aims to scale up the nuclear norm in an online manner, where the memory cost is independent of the sample size.

**Technical Challenges.** Generally speaking, the LRR problem we investigate here can be viewed as a coding algorithm, with the dataset itself being the dictionary and the low-rank matrix $X$ being the code. Using the dataset as the dictionary makes LRR difficult for online implementation for two reasons: 1) memory issues. Compared to sparse coding [25] – another popular coding algorithm whose dictionary is with fixed size and hence can be kept in memory, the dictionary of LRR grows with samples. 2) More seriously, at each iteration, we can only access a part of the dictionary (in fact, only one column of $Z$), and we aim to recover the corresponding entries in $X$ while preserving a global low-rank structure for it. This makes our problem setting totally different from [25], although part of our proof is inspired by theirs.

Our main contributions are three-folds: 1) We study the low-rank subspace clustering problem as an example and propose to scale up the nuclear norm in an online manner, which reduces the memory cost from $O(n^2)$ to $O(pd)$, where $d$ is some expected matrix rank. 2) By introducing an auxiliary variable, our algorithm can simultaneously recover the underlying subspace, estimate the representation matrix $X$ and identify the corruption $E$. 3) We theoretically show that the sequence of solutions produced by our algorithm converge to a stationary point of LRR formulation asymptotically.

### 1.1 Related Work

We discuss some related work in the literature. Some of the previous work concerns the theoretical analysis on the conditions under which the nuclear norm based program can exactly recover a low rank matrix. As the matrix nuclear norm is analogous to the vector $\ell_1$ norm, [2] extended the consistency results of Lasso to the nuclear norm minimization problem and provide sufficient and necessary conditions for the rank consistency. [9] concerned the nuclear norm regularized matrix completion problem and showed that under some mild assumptions, the missing entries can be exactly recovered. [11] and [8] then independently proved that under similar conditions, solving a convex program consisting of a nuclear norm regularization and a weighted $\ell_1$ norm penalty can exactly recover the entries that may be arbitrarily (but sparsely) corrupted. [28] showed that a family of rank minimization problems with affine constraints can be exactly solved when the mapping operator satisfies some restricted isometry property [10]. Recently, [29] tackled the unrealistic assumptions of previous work by adding some very mild assumptions, which bridges the gap between the theory and practice.

There is also a considerable amount of work on the efficient implementation of solving nuclear norm regularized problems. [7] proposed the singular value thresholding (SVT) technique and gave convergence analysis. However, since SVT calls SVD in each iteration, it is not scalable to large problems. [19] utilized a sparse SDP solver developed by [16] to derive a simple yet efficient algorithm. However, the memory requirement of their algorithm is proportional to the observed entries, making it impractical when the regularized matrix is large and dense. [11] combined stochastic subgradient and incremental SVD to boost efficiency. But for the LRR problem, the type of the loss function does not meet their requirements and thus, it is still computationally and memory expensive to use their algorithm for LRR. A recently proposed algorithm in [17] employed an active subspace selection strategy for efficiency. However, it also required $O(n^2)$ memory space, which is not practical when $n$ is large.

To tackle the memory bottleneck, [14] utilized a reformulation of the nuclear norm and proposed an online algorithm for RPCA. However, their algorithm cannot address the partial dictionary issue that emerges in online implementation of LRR. Moreover, while they proposed to simultaneously recover the subspace and identify the corruption, our algorithm can additionally learn the clustering matrix.
2 Notations and Problem Setup

We use bold letters, e.g., $\mathbf{v}$ to denote a vector. The $\ell_2$ norm and $\ell_1$ norm of a vector $\mathbf{v}$ are denoted by $\|\mathbf{v}\|_2$ and $\|\mathbf{v}\|_1$ respectively. Capital letters such as $M$ are used to denote a matrix, and its transpose is denoted by $M^\top$. The capital letter $I$ is reserved for identity matrix. The $j$-th column of a matrix $M$ is denoted by $m_j$. There are three matrix norms that will be used: $\|M\|_*$ for the nuclear norm, $\|M\|_F$ for the Frobenius norm and $\|M\|_1$ is the $\ell_1$ norm of the matrix seen as a long vector. For an integer $n$, we use $[n]$ to denote the integer set $\{1, 2, \cdots, n\}$. Finally, the trace of a square matrix $M$ is denoted as $\text{Tr}(M)$.

Our goal is to learn the representation matrix $X$ and corruption matrix $E$ in an online manner. The first technique for our online low-rank subspace clustering (OLRSC) algorithm is a non-convex reformulation of the nuclear norm. Assume the rank of $X$ is at most $d$. Then \cite{28} shows that,

$$\|X\|_* = \min_{U, V, X = UV^\top} \frac{1}{2} \left( \|U\|_F^2 + \|V\|_F^2 \right), \quad (2.1)$$

where $U \in \mathbb{R}^{n \times d}$ and $V \in \mathbb{R}^{n \times d}$. In this way, Eq. (1.1) can be written as:

$$\min_{U, V, E} \frac{1}{2} \left( \|U\|_F^2 + \|V\|_F^2 \right) + \lambda \|E\|_1,$$

s.t. $Z = ZU^\top + E$.

Note that this technique is also utilized in \cite{14} for online RPCA. However, it is unfortunate that the size of $U$ and $V$ in our problem are both proportional to $n$ and the dictionary is partially observed in each iteration, which results inapplicability of their algorithm to LRR. Another challenge for online implementation is that, all the rows of $U$ are coupled together at this moment as $U$ is left multiplied by $Z$ in the constraint. This makes it hard to sequentially update the rows of $U$ (each row of $U$ corresponds to a row in $X$).

To handle these two issues, our main novelty is to introduce an auxiliary variable $D = ZU$, whose size is $p$ by $d$ that is independent from the sample size $n$. Note that in this way, we have $Z = DV^\top + E$, which provides an intuition on the role of $D$. Namely, $D$ can be seen as a basis dictionary of the clean data, with each row of $V$ being the coefficient for each sample.

Combining these two key techniques, we derive an equivalent reformulation of LRR:

$$\min_{U, V, E} \frac{1}{2} \left( \|U\|_F^2 + \|V\|_F^2 \right) + \lambda \|E\|_1,$$

s.t. $Z = DV^\top + E$, $\quad D = ZU$.

By penalizing the equality constraints in the objective function, we obtain a regularized version of the above formulation as follows:

$$\min_{D, U, V, E} \frac{\lambda_1}{2} \|Z - DV^\top - E\|_F^2 + \frac{1}{2} \left( \|U\|_F^2 + \|V\|_F^2 \right)$$

$$\quad + \frac{\lambda_2}{2} \|D - ZU\|_F^2, \quad (2.2)$$

Let $z_i, \ e_i, \ u_i, \ v_i$ be the $i$-th column of matrices $Z, \ E, \ U^\top$ and $V^\top$ respectively and define the following two functions:

$$\tilde{\ell}(z, D, v, e) = \frac{\lambda_1}{2} \|z - Dv - e\|_2^2 + \frac{1}{2} \|v\|_2^2 + \lambda_2 \|e\|_1, \quad (2.3)$$

and

$$\tilde{\ell}(Z, D, U) = \sum_{i=1}^n \frac{1}{2} \|u_i\|_2^2 + \frac{\lambda_2}{2} \|D - \sum_{i=1}^n z_i u_i^\top\|_F^2. \quad (2.4)$$

Then Eq. (2.2) can be rewritten as:

$$\min_{D} \min_{U, V, E} \sum_{i=1}^n \tilde{\ell}(z_i, D, v_i, e_i) + \tilde{\ell}(Z, D, U), \quad (2.5)$$
which is equivalent to minimizing the following empirical loss function:

\[ f_n(D) \triangleq \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, D) + \frac{1}{n} h(Z, D), \quad (2.6) \]

where

\[ \ell(z, D) \triangleq \min_{v, e} \ell(z, D, v, e), \]

and

\[ h(Z, D) \triangleq \min_U h(Z, D, U). \quad (2.8) \]

In Section 3 we will present our online algorithm for Problem (2.6). Before that, we shall first derive the optimal solutions \( U^*, V^* \) and \( E^* \) for Eq. (2.6), based on which we are able to formulate the expected loss function.

### 2.1 Optimal \( U^*, V^* \) and \( E^* \) of Eq. (2.6)

Given \( D \), we need to compute the optimal solutions \( U^*, V^* \) and \( E^* \) to evaluate the objective value of \( f_n(D) \). What is of interest here is that the optimization procedure of \( U \) is totally different from \( V \) and \( E \). According to Eq. (2.7), when \( D \) is given, each \( v_i^* \) and \( e_i^* \) can be solved by only accessing the \( i \)-th sample \( z_i \). However, the optimal \( u_i^* \) depends on the whole dataset as the second term in Eq. (2.4) couples all the \( u_i \)’s. Fortunately, we can obtain a closed form solution for the \( u_i \)’s as follows:

\[ u_i^* = \frac{1}{n} D^\top \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i, \quad \forall i \in [n], \quad (2.9) \]

where \( N_n = \sum_{i=1}^{n} z_i z_i^\top \). Thus,

\[ h(Z, D) = \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{2} \left\| D^\top \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i \right\|_2^2 + \frac{\lambda_3}{2n^2} \left\| \left( \frac{1}{n} I + \frac{\lambda_3}{n} N_n \right)^{-1} D \right\|_F^2. \]

### 2.2 Expected Loss Function

In stochastic optimization, we are mainly interested in the expected loss function, which is defined as the limit of the empirical loss function when \( n \) tends to infinity. If we assume that all the samples are drawn i.i.d. from some unknown distribution, we have

\[ \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \ell(z_i, D) = \mathbb{E}_z[\ell(z, D)]. \]

If we further assume that the smallest singular value of \( \frac{1}{n} N_n \) is bounded away from zero, we have

\[ \lim_{n \to \infty} \frac{1}{n} h(Z, D) = \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{2} \left\| D^\top \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i \right\|_2^2 \leq \lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} c = 0. \]

Here \( c \) is some constant. Also it is clear that \( h(Z, D) \) is always non-negative which follows that

\[ \lim_{n \to \infty} \frac{1}{n} h(Z, D) = 0. \]

Finally, the expected loss function is formulated as follows:

\[ f(D) \triangleq \lim_{n \to \infty} f_n(D) = \mathbb{E}_z[\ell(z, D)], \quad (2.10) \]

### 3 Algorithm

Our OLRSC algorithm is summarized in Algorithm 1. We optimize the variables in an alternative manner which is shown to be a successful strategy [20]. To start up, we first initialize \( D \) with a random matrix \( D_0 \). Then at the \( t \)-th iteration, the optimal solutions \( \{v_t, e_t\} \) are produced by
Algorithm 1 OLRSC – Alternating Minimization

**Input:** \( Z \in \mathbb{R}^{p \times n} \) (observed samples), parameters \( \lambda_1, \lambda_2 \) and \( \lambda_3 \), random matrix \( D_0 \in \mathbb{R}^{p \times d} \) (initial basis), zero matrix \( M_0, A_0 \) and \( B_0 \).

**Output:** Optimal basis \( D_n \).

1: for \( t = 1 \) to \( n \) do
2: Access the \( t \)-th sample \( z_t \).
3: Compute the coefficient and noise:
   \[
   \{ v_t, e_t \} = \arg\min_{v, e} \tilde{\ell}(z_t, D_{t-1}, v, e), \quad (3.1)
   \]
   \[
   u_t = \arg\min_u \tilde{\ell}_2(z_t, D_{t-1}, M_{t-1}, u). \quad (3.2)
   \]
4: Update the accumulation matrices:
   \[
   M_t \leftarrow M_{t-1} + z_t u_t^\top,
   A_t \leftarrow A_{t-1} + v_t v_t^\top,
   B_t \leftarrow B_{t-1} + (z_t - e_t)v_t^\top.
   \]
5: Update the basis:
   \[
   D_t = \arg\min_D g_t(D)
   = \arg\min_D \left[ \frac{1}{t} \left( \frac{1}{2} \text{Tr}(D^\top D(\lambda_1 A_t + \lambda_3 I)) - \text{Tr}(D^\top (\lambda_1 B_t + \lambda_3 M_t)) \right) \right]. \quad (3.3)
   \]
6: end for

minimizing the objective function \( \tilde{\ell}(z_t, D_{t-1}, v, e) \) over \( v \) and \( e \). To obtain \( u_t \), we minimize the function \( \tilde{\ell}_2(z_t, D_{t-1}, M_{t-1}, u) \), which is defined as:

\[
\tilde{\ell}_2(z, D, M, u) = \frac{1}{2} \|u\|^2_2 + \frac{\lambda_3}{2} \|D - M u^\top\|^2_F. \quad (3.4)
\]

Finally, with all the past information \( F_t = \{ z_i, v_i, e_i, u_i \}_{i=1}^t \) on hand, we optimize the following surrogate function

\[
g_t(D) \triangleq \frac{1}{t} \left( \sum_{i=1}^t \tilde{\ell}(z_i, D, v_i, e_i) + \sum_{i=1}^t \frac{1}{2} \|u_i\|^2_2 + \frac{\lambda_3}{2} \|D - M_i\|^2_F \right) \quad (3.5)
\]
to obtain a new iterate \( D_t \).

We now explain implementation details for solving Eq. (3.1), (3.2) and (3.3) in the following subsections.

3.1 Solve Eq. (3.1)

Given \( D \), it is not easy to jointly optimize over \( v \) and \( e \). However, we note that if \( e \) is fixed, we can optimize \( v \) in closed form:

\[
v = (D^\top D + \frac{1}{\lambda_1} I)^{-1} D^\top (z - e). \quad (3.6)
\]

Conversely, given \( v \), we obtain the optimal \( e \) via soft-thresholding [15]:

\[
e = S_{\lambda_2/\lambda_1}[z - Dv]. \quad (3.7)
\]

Thus, we utilize coordinate descent algorithm to optimize \( v \) and \( e \). See Algorithm 2 in the supplement for details. Since the objective function is jointly and strongly convex over \( v \) and \( e \), we attain a global optimum when the the sequence of the iterates converges (see Section 2.7 in [3]).
3.2 Solve Eq. (3.2)

According to the first order optimality condition of Eq. (3.4), we can solve $u_t$ in closed form:

$$u_t = (\|z_t\|_2^2 + \frac{1}{\lambda_3})^{-1}(D_{t-1} - M_{t-1})^T z_t. \quad (3.8)$$

As we discussed in Section 2.1, the optimal values of variables $v$ and $e$ can be “accurately” solved by only accessing the $t$-th sample $z_t$ if $D$ is given. However, this does not hold for the variable $u$. In fact, even though $D$ is given, the optimal $u_t$ depends on the whole dataset, see Eq. (2.9). In stochastic optimization, we only store the current sample. Thus we have to “approximately” solve $u_t$.

Here, we give some intuition on the objective function $\tilde{\ell}_2^2(z, D, M, u)$. Actually, our strategy can be seen as a one-pass block coordinate descent algorithm for the objective function $\tilde{h}(Z, D, U)$. Assume we have $n$ samples in total. We initialize all the $u_t$’s with a zero vector. After accessing the $t$-th sample $z_t$, we only update $u_t$ with other $u$’s being kept. In this way, optimizing $\tilde{h}(Z, D, U)$ is equivalent to minimizing the function $\tilde{\ell}_2^2(z_t, D, M_{t-1}, u)$. So after revealing all the samples, each $u_t$ is sequentially updated only once. In this point of view, we verify that $g_t(D)$ is a surrogate function of $f_t(D)$.

3.3 Solve Eq. (3.3)

We can compute the minimizer of $g_t(D)$ by examining its first order optimality condition. In this way, we will obtain a closed form solution for $D_t$. However, for computational efficiency, we apply block coordinate descent algorithm to optimize $D$. See Algorithm 3 in the supplement for details. Also note that although we aim to minimize Eq. (3.5), which seems to require all the past information, we actually only need to record $A_t$, $B_t$ and $M_t$ as shown in Algorithm 1 whose sizes are independent of $n$.

4 Main Results and Proof Sketch

4.1 Assumptions

We make the following assumptions throughout this section.

1. The observed data are generated i.i.d. from some distribution and there exist constants $\alpha_0$ and $\alpha_1$, such that $0 < \alpha_0 \leq \|z\|_2 \leq \alpha_1$ almost surely.

2. The smallest singular value of the averaged cumulation matrix $N_t = \frac{1}{t} \sum_{i=1}^t z_i z_i^T$ is lower bounded away from zero.

3. The surrogate functions $g_t(D)$ in Eq. (3.5) are strongly convex. That is, 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 $\beta_0$. Note that we can easily enforce this assumption by adding a term $\frac{\beta_0}{t} ||D||_F^2$ to $g_t(D)$.

4.2 Main Theorem

Based on these assumptions, we establish the main result of this paper.

**Theorem 4.1 (Convergence of $D_t$).** Let $\{D_t\}$ be the optimal basis produced by Algorithm 1 and let $f(D)$ be the expected loss function defined in Eq. (2.10). Then $D_t$ converges to a stationary point of $f(D)$ when $t$ goes to infinity.

Note that since the reformulation of the nuclear norm (2.1) is non-convex, we can only guarantee that the solution is a stationary point in general [3]. However, as we will demonstrate in Section 6, the obtained solutions always work well compared to state-of-the-art solvers.

---

1Note that “accurately” and “approximately” here mean that when only $D$ and $z_t$ are given, whether we can obtain the same solution $\{v_t, e_t, u_t\}$ as for the batch problem (2.6).
4.3 Proof Outline

We sketch the proof and leave the details in the supplement. In this section, the letter $c$ and its subscript variants are reserved to denote universal constants. We first show that the positive stochastic process $w_t \overset{\Delta}{=} g_t(D_t)$ is a quasi-martingale, which is guaranteed to converge almost surely \[5\]. To this end, we need to uniformly bound the stochastic variables in our problem which is a fundamental result that will be heavily used in the subsequent proofs. That is,

**Proposition 4.2.** Let $\{u_t\}, \{v_t\}, \{e_t\}$ and $\{D_t\}$ be the optimal solutions produced by Algorithm\[7\]

Then,

1. $v_t$, $e_t$, $\frac{1}{t} A_t$ and $\frac{1}{t} B_t$ are uniformly bounded.
2. $M_t$ is uniformly bounded.
3. $D_t$ is supported by some compact set $\mathcal{D}$.
4. $u_t$ is uniformly bounded.

Next, we examine the difference of $w_{t+1}$ and $w_t$:

\[
 w_{t+1} - w_t = g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{f'_i(D_{t+1}) - g'_i(D_t)}{t+1} + \frac{\ell(z_{t+1}, D_t) - f'_i(D_t)}{t+1} + r_t, \tag{4.1}
\]

where we define $f'_i(D_t) = \frac{1}{t} \sum_{i=1}^{t} \ell(z_i, D_t)$, $g'_i(D_t) = \frac{1}{t} \sum_{i=1}^{t} \ell(z_i, D_t, v_i, e_i)$ and

\[
 r_t = \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_i}{2(t+1)} \|D_t - M_t\|^2_F - \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 - \frac{\lambda_i}{2t} \|D_t - M_t\|^2_F.
\]

Since $D_{t+1}$ minimizes $g_{t+1}(D)$, we have $g_{t+1}(D_{t+1}) - g_{t+1}(D_t) \leq 0$. Also $g'_i(D_t)$ is the surrogate of $f'_i(D_t)$, so $f'_i(D_t) - g'_i(D_t) \leq 0$. In addition, according to the definition of $M_t$ (see Algorithm\[1\]) and the closed form solution of $u_t$ \(5.8\), we can show that $r_t \leq 0$. Hence, $w_{t+1} - w_t$ is upper bounded by $\frac{\ell(z_{t+1}, D_t) - f'_i(D_t)}{t+1}$.

Then, we show that $\{\ell(z, D)\}$ is P-Donsker \[31\], for which the difference of the empirical loss and expected loss can be uniformly upper bounded by $O(1/\sqrt{t})$. Therefore, by taking the expectation conditioned on the past filtration $\mathcal{F}_t$, we can bound the conditional expectation of $w_{t+1} - w_t$ as follows:

\[
 \mathbb{E}[w_{t+1} - w_t \mid \mathcal{F}_t] \leq \frac{c}{\sqrt{t(t+1)}}.
\]

Now we have verified all the hypotheses in Lemma\[13.3\] and we conclude that $g_t(D_t)$ converges almost surely, which is formally stated in the following theorem.

**Theorem 4.3** (Convergence of the surrogate function $g_t(D_t)$). The surrogate function $g_t(D_t)$ we defined in Eq. \(5.5\) converges almost surely, where $D_t$ is the solution produced by Algorithm\[7\].

**Remark 4.1.** Note that when we bound $w_{t+1} - w_t$, a critical step is that we separate the surrogate function $g_t(D)$ into two parts: one associated with the summation of $\ell(z_i, D, v_i, e_i)$ (denoted by $g'_i(D)$) and the other associated with $u_i$’s. The first component is a surrogate of a family of P-Donsker functions, and the second component has a closed form solution. These two characterizations together allow us to derive a uniform bound for $w_{t+1} - w_t$.

Since $g_t(D)$ works as a surrogate for $f_t(D)$, we expect that $f_t(D)$ and $g_t(D)$ will converge to the same limit almost surely. The following theorem conforms our intuition.

**Theorem 4.4** (Convergence of $f_t(D_t)$). Let $f_t(D_t)$ be the empirical loss function defined in Eq. \(2.6\) and $D_t$ be the solution produced by the Algorithm\[7\] Let $b_t = g_t(D_t) - f_t(D_t)$. Then, $b_t$ converges almost surely to 0. Thus, $f_t(D_t)$ converges almost surely to the same limit of $g_t(D_t)$.

To prove Theorem\[4.4\] we again separate $f_t(D_t)$ and $g_t(D_t)$ into two components as aforementioned. Also note that $f'_t(D_t) - g'_t(D_t)$ can be expressed by the other terms in Eq. \(4.1\). These together give

\[
 \frac{b_t}{t+1} = g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{\ell(z_{t+1}, D_t) - f'_t(D_t)}{t+1} + w_t - w_{t+1} + \frac{q_t}{t+1} + r_t,
\]
where
\[
q_t = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \| u_i \|_2^2 + \frac{\lambda_3}{2t} \| D_t - M_t \|_F^2 - \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_t \right)^{-1} z_i \|^2_2 \\
- \frac{\lambda_3}{2t^3} \left( \left( I + \frac{1}{t} N_t \right)^{-1} D_t \right)_{i,j}^2.
\]

According to the uniform boundedness result from Proposition 4.2, combined with a tighter bound on \( r_t \), we prove that \( \frac{d}{dt} q_t + r_t \) can be bounded by \( \frac{1}{t^2} \). Also, taking a conditional expectation gives a bound on \( \frac{1}{t^2} \). Finally, \( w_t - w_{t+1} \) can be bounded by utilizing the result from Lemma D.3. Putting them together, we have \( \sum_{i=1}^{\infty} \frac{b_i}{t+1} < +\infty \).

Next, we prove that \( |b_{t+1} - b_t| \) can be upper bounded by \( O(1/t) \). This is verified by showing that it can be bounded by \( |g_t(D_{t+1}) - g_t(D_t)| + |f_t(D_{t+1}) - f_t(D_t)| \) plus an \( O(1/t) \) term. In addition, by examining the gradient of \( g_t(D) \) and \( f_t(D) \), we show that they are uniformly Lipschitz. Moreover, the strong convexity assumption of \( g_t(D) \) allows us to show that \( \| D_{t+1} - D_t \|_F = O(1/t) \). Combining this fact with the Lipschitz property, we establish the \( O(1/t) \) bound for \( |b_{t+1} - b_t| \).

Finally, applying Lemma D.4 we conclude that \( b_i \) converges to 0 almost surely. Namely, \( f_t(D_t) \) converges to the same limit of \( g_t(D_t) \) almost surely. Thus establishing Theorem 4.4.

Note that \( f(D) \) is the limit of \( f_t(D) \). From the central limit theorem, we know that they converge to the same limit almost surely. This simple fact gives the following theorem.

**Theorem 4.5 (Convergence of \( f(D_t) \)).** Let \( f(D) \) be the expected loss function we defined in Eq. (2.10) and let \( D_t \) be the optimal solution produced by Algorithm 7. Then \( f(D_t) \) converges almost surely to the same limit of \( g_t(D_t) \).

To prove our main theorem 4.1, we only need to verify that the gradient of \( f(D_t) \) vanishes as \( t \) goes to infinity. This can be achieved by taking a first order Taylor expansion for \( f(D_t) \) and \( g_t(D_t) \). However, it is notable that since we are now considering the infinite situation, it is essential for us to show that the second order derivatives of them always exist (or, uniformly bounded), which is equivalent to proving that \( \nabla f(D) \) and \( \nabla g_t(D) \) are uniformly Lipschitz. Equipped with this property, we show that \( \nabla f(D_t) \) goes to be equal to \( \nabla g_t(D_t) \) when \( t \) tends to infinity. Since \( D_t \) minimizes \( g_t(D) \), we always have \( \nabla g_t(D_t) = 0 \), which concludes the proof.

## 5 Extension to Matrix Completion

The vanilla matrix completion problem [9] seeks a low-rank matrix \( X \) to approximate the observed data \( Z \), subject to an equality constraint. However, it is well known that the assumption in their work that the locations of observed entries are sampled uniformly is not practical [29]. To handle the issue that the observed entries in \( Z \) are typically not uniform in real world applications, [22] proposed the following formulation (termed LRFD):

\[
\min_X \| X \|_*, \text{ s.t. } \mathcal{P}_\Omega(Z - AX) = 0,
\]

where \( A \) is a given matrix with size \( p \times n \). Although the above formulation is proved to be able to handle the non-uniform data and capture additional structure of the data, the regularized matrix \( X \) here is not necessary to be sparse and its size is \( n \times n \). Hence recently established solvers [19][18][17] are not scalable for LRFD if \( n \) is large. We show that this problem can be cast into our framework. To see this, we first construct an indicator matrix \( M \), with the \((i,j)\)-th entry of \( M \) being \( M_{ij} = 1 \) if \((i,j) \in \Omega \) and \( M_{ij} = \frac{1}{m} \) otherwise. Then one can simply verify that LRFD is equivalent to the following problem

\[
\min_{X,E} \frac{\lambda_1}{2} \| Z - AX - E \|_F^2 + \| X \|_* + \| M \circ E \|_1
\]

when \( m \) goes to infinity. Here, “\( \circ \)” denotes entry-wise product. We elaborate in the supplement that how this program can be solved by slightly modifying our algorithm.
In this section, we report some simulation results to show the effectiveness and robustness of our algorithm. We first demonstrate that OLRSC can be used for robust subspace recovery. Then we utilize spectral clustering technique \[27\] for subspace clustering.

### Data Generation
We use 5 disjoint subspaces \( \{S_k\}_{k=1}^{5} \subset \mathbb{R}^p \), whose bases are denoted by \( \{L_k\}_{k=1}^{5} \subset \mathbb{R}^{p \times d_k} \). The clean data matrix \( \bar{Z}_k \subset S_k \) is then produced by \( \bar{Z}_k = L_k R_k \), where \( R_k \subset \mathbb{R}^{n_k \times d_k} \). Both of the entries of \( L_k \)’s and \( R_k \)’s are sampled i.i.d. from the normal distribution. Finally, the observed data matrix \( Z \) is generated by \( Z = \bar{Z} + E \), where \( \bar{Z} \) is the column-wise concatenation of \( \bar{Z}_k \)’s followed by a random permutation, \( E \) is the sparse corruption whose \( \rho \) fraction entries are non-zero and follow an i.i.d. uniform distribution over \([-2, 2]\).

### Evaluation Metric
We evaluate the fitness of the recovered subspaces \( D \) and the groundtruth bases \( L \) (concatenation of \( L_k \)’s) by the Expressed Variance (EV) \[32\]. The values of EV are scaled between 0 and 1, and a higher value means better recovery. We also measure the performance of subspace clustering by Subspace Misclassification Rates (SMR) \[12\]. SMR also ranges in \([0, 1]\), and a lower value indicates a more accurate clustering.

### Other Settings
The ambient dimension \( p \) is set to be 200 and the intrinsic dimension \( d_k \)’s are set to be a same value and we use \( d = 5d_k \). Throughout our experiments, \( \lambda_1 = 1 \), \( \lambda_2 = \frac{1}{\sqrt{n}} \) and \( \lambda_3 = 1 \). We follow the default setting of the parameters for all the baselines. Each experiment will be repeated 10 times and we plot the averaged result. If not specified, there are 5000 samples in each small subspace \( S_k \), so that the total number of samples is 25,000.

### 6.1 Subspace Recovery
We first illustrate that OLRSC can effectively recover the underlying subspaces. Here, we set \( d_k = 0.05 \). We plot the EV curve against the number of samples of our algorithm and three baselines: OR-PCA \[14\], LRR and PCP \[21\]. As shown in Figure 1 LRR and PCP always achieve the best performance as they are formulated in a batch manner. Our algorithm and OR-PCA can always converge to a favorable point after revealing all the samples. However, compared with OR-PCA, OLRSC is more robust when a large fraction of entries are corrupted (\( \rho = 0.5 \)). This suggests that OLRSC can handle much harder tasks than OR-PCA possibly because OLRSC explicitly take into account that the data are generated from a union of small subspaces.

### 6.2 Subspace Clustering
We now examine the subspace clustering problem and compare our algorithm with three baselines: OR-PCA, LRR and SSC \[12\]. In order to perform clustering, we collect all the \( u \)’s and \( v \)’s to form the representation matrix \( X = UV^\top \). For OR-PCA, we use \( V_0V_0^\top \) as the similarity matrix \[23\], where \( V_0 \) is the row space of \( Z_0 = U_0 \Sigma_0 V_0^\top \) and \( Z_0 \) is the clean matrix recovered by OR-PCA. OLRSC and OR-PCA are conducted with totally 10 epochs and we plot the SMR at the end of each epoch. The reason we repeat them for several epochs is that, since each sample is only revealed once in an online algorithm, although it is sufficient to estimate the subspace, no forward correction is applied on the coefficients (e.g., \( \hat{U} \) and \( V \) in OLRSC), which will be used to cluster samples.

The subspace clustering results are illustrated in Figure 2. We observe that when the data is nearly clean (\( \rho = 0.01 \)), all the algorithms perform comparably. As we tune the \( \rho \) to be 0.2, LRR and
notice that if we pick larger selections. The relaxed requirement of consistently identify the cluster of the data points if \( O \) space clustering problem as an example, and developed an online algorithm whose memory cost is when the size of the regularized matrix is proportional to \( O(n^2) \). We considered the low-rank subspace clustering problem as an example, and developed an online algorithm whose memory cost is \( O(pd) \) – two orders of magnitude more memory efficient than the batch counterpart. Moreover, our

**Figure 2:** SMR in each epoch. A lower SMR means better subspace clustering.

**Figure 3:** The influence of \( d \) on subspace recovery.

SSC degrade a bit. When \( \rho \) is increased to 0.4 or higher, LRR and SSC cannot identify the correct cluster as they have no exact recovery guarantee for sparse corruption. However, for these hard cases, OLRSC still reports satisfactory results. Particularly, when \( \rho = 0.6 \), OLRSC can identify 10% samples more than OR-PCA after 10 epochs. We also note that for the relatively simple cases, OLRSC converges slightly slower than OR-PCA. This is possibly due to two reasons: First, we actually implement an online version for Eq. (2.2), which can be seen as a regularized version of LRR. When the data are highly corrupted, a regularized version might be a better choice. Second, the improvement of SMR may come from the non-convex formulation, which is recently shown to be superior to the convex counterpart [26]. To fully understand the rational behind this phenomenon is an important direction for future research.

### 6.3 Influence of the Rank

A key ingredient of our formulation is a factorization on the regularized matrix, which requires an explicit estimation on the rank of the clean data \( \bar{Z} \). Here we examine the influence of the selection of \( d \) (which plays as an upper bound of the true rank). We report both EV and SMR for different \( d \) under a range of corruptions. We set the number of samples in each subspace to be 1000 and \( d_k = 10 \). Since the five subspaces are disjoint, the true rank is 50. We also repeat the optimization procedure for 10 epochs.

From Figure 4 we observe that our algorithm cannot recover the true subspace if \( d \) is smaller than the true rank. On the other hand, when \( d \) is sufficiently large (at least larger than the true rank), our algorithm can perfectly estimate the subspace. This agrees with the results in [6].

In Figure 4 we illustrate the influence of \( d \) on subspace clustering. Generally speaking, OLRSC can consistently identify the cluster of the data points if \( d \) is sufficiently large. However, different from the subspace recovery task, here the requirement for \( d \) seems to be slightly relaxed. In particular, we notice that if we pick \( d \) as 40 (smaller than the true rank), OLRSC still performs comparably to the larger selections. The relaxed requirement of \( d \) may benefit from the fact that the spectral clustering step in the subspace clustering task can correct some wrong points to a degree as suggested by [30].

### 7 Conclusion

In this paper, we proposed a method to scale up the nuclear norm in a memory efficient manner when the size of the regularized matrix is proportional to \( O(n^2) \). We considered the low-rank subspace clustering problem as an example, and developed an online algorithm whose memory cost is \( O(pd) \) – two orders of magnitude more memory efficient than the batch counterpart. Moreover, our
algorithm can simultaneously recover the underlying subspaces and learn the clustering matrix, as well as identify the possible corruptions. The two key techniques of our algorithm are to utilize a non-convex reformulation of the nuclear norm and to introduce an auxiliary variable. We then presented an alternating minimization schema to optimize the stochastic variables. We also established a theoretical guarantee that the solution produced by our algorithm will converge to a stationary point. Moreover, we extended our algorithm to other significant problems such as matrix completion. Finally, via simulations, we empirically showed the robustness and effectiveness of OLRSC. Notably, compared with a recently developed online algorithm OR-PCA, OLRSC is more robust for subspace recovery and subspace clustering when a large number of entries are corrupted.
A Optimal solution $U^*$ for Eq. (2.8)

The optimal solution $U$ for Eq. (2.8) is given by:

$$\frac{\partial \tilde{h}(Z, D, U)}{\partial U} = U + \lambda_3 (Z^T ZU - Z^T D) = 0$$

$$\Rightarrow U^* = (\frac{1}{\lambda_3} I + Z^T Z)^{-1} Z^T D$$

Then,

$$U^*^T = D^T (\frac{1}{\lambda_3} I + Z^T Z)^{-1}$$

$$= \lambda_3 D^T Z \sum_{j=0}^{\infty} (-\lambda_3 Z^T Z)^j$$

$$= \lambda_3 D^T \left[ \sum_{j=0}^{\infty} (-\lambda_3 ZZ^T)^j \right] Z$$

$$= D^T \left( \frac{1}{\lambda_3} I + ZZ^T \right)^{-1} Z$$

$$= D^T \left( \frac{1}{\lambda_3} I + \sum_{i=1}^{n} z_i z_i^T \right)^{-1} Z$$

Note that $u_i$ is the column of $U^T$. So for each $i \in [n]$,

$$u_i^* = D^T \left( \frac{1}{\lambda_3} I + \sum_{i=1}^{n} z_i z_i^T \right)^{-1} z_i$$

$$= \frac{1}{n} D^T \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i$$

Here, $N_n = \sum_{i=1}^{n} z_i z_i^T$.

Also,

$$ZU^*^T = Z \left( \frac{1}{\lambda_3} I + Z^T Z \right)^{-1} Z^T D$$

$$= \lambda_3 Z \left( I + \lambda_3 Z^T Z \right)^{-1} Z^T D$$

$$= \lambda_3 \left[ \sum_{j=0}^{\infty} (-\lambda_3 Z^T Z)^j \right] Z^T D$$

$$= \lambda_3 \sum_{j=0}^{\infty} (-\lambda_3)^j (ZZ^T)^{j+1} D$$

$$= D - \left( I + \lambda_3 ZZ^T \right)^{-1} D$$

$$= D - \left( I + \lambda_3 \sum_{i=1}^{n} z_i z_i^T \right)^{-1} D$$

$$= D - \frac{1}{n} \left( \frac{1}{n} I + \frac{\lambda_3}{n} N_n \right)^{-1} D$$
Thus,
\[
h(Z, D) = \sum_{i=1}^{n} \frac{1}{2} \left\| \frac{1}{n} D^\top \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i \right\|^2_2 \\
+ \frac{\lambda_3}{2} \left\| \frac{1}{n} \left( \frac{1}{n} I + \frac{1}{n} N_n \right)^{-1} D \right\|_F^2 \\
= \frac{1}{n^2} \sum_{i=1}^{n} \frac{1}{2} \left\| D^\top \left( \frac{1}{\lambda_3 n} I + \frac{1}{n} N_n \right)^{-1} z_i \right\|^2_2 \\
+ \frac{\lambda_3}{2n^2} \left\| \left( \frac{1}{n} I + \frac{1}{n} N_n \right)^{-1} D \right\|_F^2.
\]

B Algorithm Details

**Algorithm 2** Solve Eq. (3.1)

**Input:** \( D \in \mathbb{R}^{p \times d} \), \( z \in \mathbb{R}^p \), parameter \( \lambda_1 \) and \( \lambda_2 \)

**Output:** Optimal \( v \) and \( e \).

1: Set \( e = 0 \).
2: repeat
3: Update \( v \):
\[
v = (D^\top D + \frac{1}{\lambda_1} I)^{-1} D^\top (z - e).
\]
4: Update the noise \( e \):
\[
e = S_{\lambda_2 / \lambda_1} [z - Dv].
\]
5: until convergence

**Algorithm 3** Solve Eq. (3.3)

**Input:** \( D \in \mathbb{R}^{p \times d} \) in the previous iteration, accumulation matrix \( M \), \( A \) and \( B \), parameter \( \lambda_1 \) and \( \lambda_3 \).

**Output:** Optimal \( D \) (updated).

1: Denote \( \tilde{A} = \lambda_1 A + \lambda_3 I \) and \( \tilde{B} = \lambda_1 B + \lambda_3 M \).
2: repeat
3: for \( j = 1 \) to \( d \) do
4: Update the \( j \)-th column of \( D \):
\[
d_j \leftarrow d_j - \frac{1}{\tilde{A}_j} \left( D \tilde{a}_j - \tilde{b}_j \right)
\]
5: end for
6: until convergence

For Algorithm 2 we set a threshold \( \epsilon = 0.001 \). Let \( \{v', e'\} \) and \( \{v'', e''\} \) be the two consecutive iterates. If the maximum of \( \|v' - v''\|_2 \) and \( \|e' - e''\|_2 \) is less than \( \epsilon \), then we stop Algorithm 2.

For Algorithm 3 we observe that a one-pass update on the dictionary \( D \) is enough for the final convergence of \( D \), as we shown in the experiments.
C Extension to Matrix Completion

For the vanilla matrix completion problem, it approximates the observed data Z with a low-rank matrix X. That is,

$$\min_X \|X\|_*, \quad \text{s.t. } P_{\Omega}(Z - X) = 0.$$ 

Here, $\Omega$ is the index set of observed entries and $P_{\Omega}(M) = m_{ij}$ if $(i, j) \in \Omega$ and zeros otherwise. In such formulation, we sometimes can expect that $X$ is sparse since $Z$ is sparse, which allows efficient computation of recent solvers [19, 1, 17].

However, as suggested by [22], in practice, samples in $Z$ will exhibit clustering structure beyond the low-rankness. To better handle such additional structure, they proposed a new formulation termed LRFD as we mentioned in Section 5:

$$\min_X \|X\|_*, \quad \text{s.t. } P_{\Omega}(Z - AX) = 0.$$ 

The benefit of LRFD is that it captures both the low-rank and clustering structure. However, at this moment, $AX$ is used to approximate $Z$ instead of $X$. So it is not necessary for $X$ to be sparse, which hinders the applicability of recently established batch based solvers.

Now we show how Eq. (5.1) can be solved by slightly modifying our OLRSC algorithm. By utilizing the factorization form (2.1) and introduce $D = AU$, we obtain:

$$\min_{U, V, E} \frac{\lambda_1}{2} \|Z - DV^T - E\|_F^2 + \frac{1}{2} \left(\|U\|_F^2 + \|V\|_F^2\right) + \|M \circ E\|_1.$$ 

Define

$$\tilde{\ell}(z, D, v, e) = \frac{\lambda_1}{2} \|z - Dv - e\|_2^2 + \frac{1}{2} \|v\|_2^2 + \|m \circ e\|_1,$$

$$\tilde{h}(Z, D, U) = \frac{1}{2} \|U\|_F^2 + \frac{\lambda_3}{2} \|D - AU\|_F^2,$$

and

$$\ell(z, D) = \min_{v, e} \tilde{\ell}(z, D, v, e),$$

$$h(Z, D) = \min_U \tilde{h}(Z, D, U).$$

One can easily check that all other derivations, such as empirical loss function, expected loss function, surrogate function, are same with OLRSC. What we need to modify is the optimization schema for $\ell(z, D)$. To be more detailed, for this problem, the optimal $e$ is given by a different way. Let

$$\Omega_1 = \{j \mid m(j) = m, j \in [p]\},$$

$$\Omega_2 = \{j \mid m(j) = \frac{1}{m}, j \in [p]\},$$

where $m(j)$ is the $j$-th element of $m$. Then, $\tilde{\ell}(z, D, v, e)$ can be rewritten as:

$$\tilde{\ell}(z, D, v, e)$$

$$= \left(\frac{\lambda_1}{2} \|z_{\Omega_1} - (Dv)_{\Omega_1} - e_{\Omega_1}\|_2^2 + \frac{1}{2} \|v_{\Omega_1}\|_2^2 + \frac{\lambda_3}{m} \|e_{\Omega_1}\|_1\right)$$

$$+ \left(\frac{\lambda_1}{2} \|z_{\Omega_2} - (Dv)_{\Omega_2} - e_{\Omega_2}\|_2^2 + \frac{1}{2} \|v_{\Omega_2}\|_2^2 + \frac{\lambda_3}{m} \|e_{\Omega_2}\|_1\right).$$

Note that $z$ and $D$ are given which allows us to compute the optimal $e_{\Omega_1}$ and $e_{\Omega_2}$ separately via soft-thresholding.
D Proof Preliminaries

Lemma D.1 (Corollary of Theorem 4.1 from [3]). Let \( f : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{R} \). Suppose that for all \( x \in \mathbb{R}^p \) the function \( f(x, \cdot) \) is differentiable, and that \( f \) and \( \nabla_u f(x,u) \) are continuous on \( \mathbb{R}^p \times \mathbb{R}^q \). Let \( v(u) \) be the optimal value function \( v(u) = \min_{x \in C} f(x,u) \), where \( C \) is a compact subset of \( \mathbb{R}^p \). Then \( v(u) \) is directionally differentiable. Furthermore, if for \( u_0 \in \mathbb{R}^q \), \( f(\cdot, u_0) \) has unique minimizer \( x_0 \) then \( v(u) \) is differentiable in \( u_0 \) and \( \nabla_u v(u_0) = \nabla_u f(x_0, u_0) \).

Lemma D.2 (Corollary of Donsker theorem [31]). Let \( F = \{ f_0 : \mathcal{X} \to \mathbb{R}, \theta \in \Theta \} \) be a set of measurable functions indexed by a bounded subset \( \Theta \) of \( \mathbb{R}^d \). Suppose that there exists a constant \( K \) such that

\[
|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq K \| \theta_1 - \theta_2 \|_2 ,
\]

for every \( \theta_1 \) and \( \theta_2 \) in \( \Theta \) 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 \( G_n f \) as

\[
\mathbb{P}_n f = \frac{1}{n} \sum_{i=1}^{n} f(X_i) ,
\]

\[
\mathbb{P} f = \mathbb{E}[f(X)] ,
\]

\[
G_n f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P} f) .
\]

Lemma D.3 (Sufficient condition of convergence for a stochastic process [5]). Let \( (\Omega, \mathcal{F}, P) \) be a measurable probability space, \( u_t \), for \( t \geq 0 \), be the realization of a stochastic process and \( \mathcal{F}_t \) be the filtration by the past information at time \( t \). Let

\[
\delta_t = 1 \text{ if } \mathbb{E}[|u_{t+1} - u_t| | \mathcal{F}_t] > 0 ,
\]

\[
0 \text{ otherwise.}
\]

If for all \( t \), \( u_t \geq 0 \) 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,

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

Lemma D.4 (Lemma 8 from [25]). Let \( a_t, b_t \) be two real sequences such that for all \( t \), \( a_t \geq 0 \), \( b_t \geq 0 \), \( \sum_{t=1}^{\infty} a_t = \infty \), \( \sum_{t=1}^{\infty} a_t b_t < \infty \), \( \exists K > 0 \), such that \( |b_{t+1} - b_t| < K a_t \). Then, \( \lim_{t \to +\infty} b_t = 0 \).

E Proof Details

E.1 Proof of Boundedness

**Proposition E.1.** Let \( \{ u_t \} \), \( \{ v_t \} \), \( \{ e_t \} \) and \( \{ D_t \} \) be the optimal solutions produced by Algorithm [7].

Then,

1. \( v_t, e_t, \frac{1}{t} A_t \) and \( \frac{1}{t} B_t \) are uniformly bounded.
2. \( M_t \) is uniformly bounded.
3. \( D_t \) is supported by some compact set \( \mathcal{D} \).
4. \( u_t \) is uniformly bounded.

**Proof.** Let us consider the optimization problem [3, 33]. As the trivial solution \( \{ v_t', e_t' \} = \{ 0, 0 \} \) are feasible, we have

\[
\tilde{\ell}_1(z_t, D_{t-1}, 0, 0) = \frac{\lambda_1}{2} \| z_t \|_2^2 .
\]
Therefore, the optimal solution should satisfy:
\[
\frac{1}{2} \| z_t - D_{t-1} v_t - e_t \|_2^2 + \frac{1}{2} \| v_t \|_2^2 + \lambda_2 \| e_t \|_1 \leq \frac{1}{2} \| z_t \|_2^2,
\]
which implies
\[
\frac{1}{2} \| v_t \|_2^2 \leq \frac{\lambda_2}{\lambda_1} \| z_t \|_2^2,
\]
\[
\lambda_2 \| e_t \|_1 \leq \frac{\lambda_2}{\lambda_1} \| z_t \|_2^2.
\]
Since \( z_t \) is uniformly bounded (Assumption 1), \( v_t \) and \( e_t \) are uniformly bounded.
To examine the uniform bound for \( \frac{1}{t} A_t \) and \( \frac{1}{t} B_t \), note that
\[
\frac{1}{t} A_t = \frac{1}{t} \sum_{i=1}^{t} v_i v_i^\top,
\]
\[
\frac{1}{t} B_t = \frac{1}{t} \sum_{i=1}^{t} (z_i - e_i) v_i^\top.
\]
Since for each \( i \), \( v_i \), \( e_i \) and \( z_i \) are uniformly bounded, \( \frac{1}{t} A_t \) and \( \frac{1}{t} B_t \) are uniformly bounded.

Now we derive the bound for \( M_t \). All the information we have is:

1. \( M_t = \sum_{i=1}^{t} z_i u_i^\top \) (definition of \( M_t \)).
2. \( u_t = (\| z_t \|_2^2 + \frac{1}{\lambda_3})^{-1} (D_{t-1} - M_{t-1})^\top z_t \) (closed form solution).
3. \( D_t (\lambda_1 A_t + \lambda_3 I) = \lambda_1 B_t + \lambda_3 M_t \) (first order optimality condition of Eq. (3.8)).
4. \( \frac{1}{t} A_t, \frac{1}{t} B_t, \frac{1}{t} N_t \) are uniformly upper bounded (Claim 1).
5. The smallest singular values of \( \frac{1}{t} N_t \) and \( \frac{1}{t} A_t \) are uniformly lower bounded away from zero (Assumption 2 and 3).

For simplicity, we write \( D_t \) as:
\[
D_t = (\lambda_1 B_t + \lambda_3 M_t) Q_t^{-1},
\]
where
\[
Q_t = \lambda_1 A_t + \lambda_3 I.
\]
Note that as we assume \( \frac{1}{t} A_t \) is positive definite, \( Q_t \) is always invertible.
From the definition of \( M_t \) and Eq. (3.8), we know that
\[
M_{t+1} - M_t = z_{t+1} u_{t+1}^\top = \left( \| z_{t+1} \|_2^2 + \frac{1}{\lambda_3} \right)^{-1} z_{t+1} z_{t+1}^\top (D_t - M_t)
\]
\[
= P_t D_t - P_t M_t
\]
\[
= P_t (\lambda_1 B_t + \lambda_3 M_t) Q_t^{-1} - P_t M_t,
\]
where
\[
P_t = \left( \| z_{t+1} \|_2^2 + \frac{1}{\lambda_3} \right)^{-1} z_{t+1} z_{t+1}^\top.
\]
By multiplying \( Q_t \) on both sides of Eq. (E.2), we have
\[
(M_{t+1} - M_t) Q_t = \lambda_1 P_t B_t - \lambda_1 P_t M_t A_t
\]
\[
\Rightarrow M_{t+1} = (M_t - \lambda_1 P_t M_t A_t Q_t^{-1}) + \lambda_1 P_t B_t Q_t^{-1}
\]
(E.3)
By applying the Taylor expansion on $Q_t^{-1}$, we have

$$Q_t^{-1} = (\lambda_1 A_t + \lambda_3 I)^{-1} = \frac{1}{\lambda_3} \sum_{i=0}^{+\infty} \left(-\frac{\lambda_1}{\lambda_3} A_t\right)^i.$$

Thus,

$$A_t Q_t^{-1} = \frac{1}{\lambda_3} \sum_{i=0}^{+\infty} \left(-\frac{\lambda_1}{\lambda_3} A_t\right)^{i+1} = \frac{1}{\lambda_1} \left[\sum_{i=-1}^{+\infty} \left(-\frac{\lambda_1}{\lambda_3} A_t\right)^{i+1} - 1\right] = \frac{1}{\lambda_1} \left(I + \frac{\lambda_1}{\lambda_3} A_t\right)^{-1} + \frac{1}{\lambda_1} I.$$

So $M_{t+1}$ is given by

$$M_{t+1} = (I - P_t) M_t + P_t M_t \left(I + \frac{\lambda_1}{\lambda_3} A_t\right)^{-1} + \lambda_3 P_t B_t Q_t^{-1}.$$  \hspace{1cm} (E.4)

We first show that $P_t B_t Q_t^{-1}$ is uniformly bounded.

$$\|P_t B_t Q_t^{-1}\| = \left\|P_t \left(\frac{1}{t} B_t\right) \left(\frac{1}{t} Q_t\right)^{-1}\right\| \leq \|P_t\| \left\|\frac{1}{t} B_t\right\| \left\|\left(\frac{1}{t} Q_t\right)^{-1}\right\|.$$

Since we assume that \{z_t\} are uniformly upper bounded (Assumption 1), there exists a constant $\alpha_1$, such that for all $t > 0$,

$$\|z_t\|_2 \leq \alpha_1.$$

So we have

$$\|P_{t+1}\| \leq \frac{\lambda_3 \alpha_1^2}{\lambda_3 \alpha_1^2 + 1}.$$

Next, as we have shown that $\frac{1}{t} B_t$ can be uniformly bounded, there exists a constant $c_1$, such that for all $t > 0$,

$$\left\|\frac{1}{t} B_t\right\| \leq c_1.$$

And,

$$\left\|\left(\frac{1}{t} Q_t\right)^{-1}\right\| = \frac{1}{\sigma_{\min}\left(\frac{1}{t} Q_t\right)} \leq \frac{1}{\sigma_{\min}\left(\frac{1}{t} A_t + \frac{\lambda_1}{t} I\right)} \leq \frac{1}{\lambda_3 + \lambda_1 \beta_0}.$$
Thus, $\lambda_1 P_t B_t Q_t^{-1}$ is uniformly bounded by a constant, say $c_2$. That is,

$$\|\lambda_1 P_t B_t Q_t^{-1}\| \leq c_2. \quad (E.5)$$

It follows that $W_t$ can be bounded:

$$\|W_t\| \leq \|P_t\| \cdot \|M_t\| \cdot \left\|\left(I + \frac{\lambda_1}{\lambda_3} A_t\right)^{-1}\right\| + c_2$$

$$\leq \frac{\zeta_1}{\alpha_1^2 + \frac{1}{\lambda_3}} \cdot \frac{\lambda_3}{\lambda_3 + \lambda_1 \beta_0} \cdot \|M_t\| + c_2 \quad (E.6)$$

where $\zeta_1$ is derived by utilizing the assumption that $z$ is upper bounded by $\alpha_1$ and the smallest singular value of $\frac{1}{t} A_t$ is lower bounded by $\beta_0$. The last inequality always holds for some uniform constant $c_3$.

From Assumption 2 we know that the singular values of $\frac{1}{t} \sum_{i=1}^t z_i z_i^\top$ should uniformly span the diagonal. Thus, there exists a constant $\tau$, such that for all $i > 0$, $\frac{1}{t} \sum_{i=1}^{i+\tau} z_i z_i^\top$ is uniformly bounded away from zero with high probability.

Let $m_1 = \|M_1\|$. Now we pick a constant $t^*$, such that

$$\frac{c_3 \tau}{t^*} \left(\frac{1}{\alpha_0} + 1\right) \leq 0.5. \quad (E.7)$$

We also have a constant $w^*$, such that for all $t \leq t^*$,

$$\|W_t\| \leq w^*,$$

$$\frac{c_3}{t} m_1 + 0.5 w^* + c_2 \leq w^*. \quad (E.8)$$

Based on this, we first derive a bound for all $\|M_t\|$, with $t \leq t^*$. We know that there exists an integer $k^*$ (which is a uniform constant), such that $k^*(\tau + 1) \leq t^* < (k^* + 1)(\tau + 1)$. Our strategy is to bound $\|M_t\|$ in each interval $[(k-1)(\tau + 1), k(\tau + 1)]$. We start our reasoning from the first interval $[1, \tau + 1]$.

It is easy to induce from Eq. (E.4) that for all $t > 0$,

$$M_{t+1} = \prod_{i=1}^t (I - P_t) M_1 + \sum_{j=1}^{t-1} \prod_{i=j+1}^t (I - P_t) W_j + W_t.$$  

Thus,

$$\|M_{t+1}\|$$

$$= \left\|\prod_{i=1}^t (I - P_t) M_1 + \sum_{j=1}^{t-1} \prod_{i=j+1}^t (I - P_t) W_j + W_\tau\right\|$$

$$\leq \left\|\prod_{i=1}^t (I - P_t) M_1\right\| + \left\|\sum_{j=1}^{t-1} \prod_{i=j+1}^t (I - P_t) W_j + W_\tau\right\|$$

$$\leq \zeta_1 \cdot \|M_t\| + \tau w^*$$

$$\leq \left(1 - \alpha_0\right) m_1 + \tau w^*.$$  

Here, $\zeta_1$ holds because $\left\|\prod_{i=j+1}^t (I - P_t)\right\| \leq 1$ for all $j \in [\tau - 1]$. $\zeta_2$ holds because the singular values of $P_t$'s have span over the diagonal so the largest singular value of $\prod_{i=1}^t (I - P_t)$ is $1 - \alpha_0$, where $\alpha_0$ is the lower bound for all $z_i$'s (see Assumption 1).
For $M_{2(\tau+1)}$, we can similarly obtain
\[
\|M_{2(\tau+1)}\| \leq (1 - \alpha_0)^2 m_1 + (1 - \alpha_0)\tau w^* + \tau w^*.
\]

More generally, for any integer $k \leq k^*$,
\[
\|M_{k(\tau+1)}\| \leq (1 - \alpha_0)^k m_1 + \sum_{j=0}^{k-1} (1 - \alpha_0)^j \tau w^*
\]
\[
\leq m_1 + \frac{\tau w^*}{\alpha_0}.
\]

Hence, we obtain a uniform bound for $\|M_{k(\tau+1)}\|$, with $k \in [k^*]$. For any $i \in ((k-1)(\tau+1), k(\tau+1))$, they can simply bounded by
\[
\|M_i\| \leq m_1 + \frac{\tau w^*}{\alpha_0} + (i - (k - 1)(\tau + 1))w^*
\]
\[
\leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*.
\]

Therefore, for all the current $M_i$’s, we can bound them as follows:
\[
\|M_i\| \leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*, \forall t = 1, 2, \ldots, t^*.
\] (E.9)

From Eq. (E.8) and Eq. (E.9), we know that for all $t \leq t^*$,
\[
\|W_t\| \leq w^*,
\]
\[
\|M_t\| \leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*.
\]

Next, we show that the bounds still hold for $\|W_{t^*+1}\|$ and $\|M_{t^*+1}\|$, which completes our induction.

For $\|M_{t^*+1}\|$, it can simply be bounded in the same way as aforementioned because all the $W_t$’s are bounded by $w^*$ for $t < t^* + 1$. That is,
\[
\|M_{t^*+1}\| \leq \|M_{k^*(\tau+1)}\| + (t^* + 1 - k^*(\tau + 1))w^*
\]
\[
\leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*.
\] (E.10)

For $\|W_{t^*+1}\|$, from Eq. (E.6), we know
\[
\|W_{t^*+1}\| \leq \frac{c_3}{t^* + 1} \|M_{t^*+1}\| + c_2
\]
\[
\leq \frac{c_3}{t^* + 1} (m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*) + c_2
\]
\[
= \frac{c_3 m_1}{t^* + 1} + \frac{c_3 \tau}{t^* + 1} + \frac{1}{\alpha_0} + 1)w^* + c_2
\]
\[
\leq \frac{c_3 m_1}{t^* + 1} + 0.5w^* + c_2
\]
\[
\leq c_2 w^*.
\] (E.11)

Here, $\zeta_1$ is derived by utilizing Eq. (E.7) and $\zeta_2$ is derived by Eq. (E.8).

From Eq. (E.10) and Eq. (E.11), we know that the bound for $\|M_t\|$ and $\|W_t\|$’s, with $t \leq t^*$, still holds for $t = t^* + 1$. Thus we complete the induction and conclude that for all $t > 0$, we have
\[
\|M_t\| \leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*,
\]
\[
\|W_t\| \leq w^*.
\]

Thus, $M_t$ is uniformly bounded.
From Eq. (E.1), we know that
\[
D_t = \lambda_1 B_t \left( \lambda_1 A_t + \lambda_3 I \right)^{-1} + \lambda_3 M_t \left( \lambda_1 A_t + \lambda_3 I \right)^{-1} \\
= \lambda_1 \left( \frac{1}{t} B_t \right) \left( \frac{\lambda_1}{t} A_t + \frac{\lambda_3}{t} I \right)^{-1} \\
+ \frac{\lambda_3}{t} M_t \left( \frac{\lambda_1}{t} A_t + \frac{\lambda_3}{t} I \right)^{-1}.
\]

Since \( \frac{1}{t} A_t, \frac{1}{t} B_t \) and \( M_t \) are all uniformly bounded, \( D_t \) is also uniformly bounded.

By examining the closed form of \( u_t \), and note that we have proved the uniform boundedness of \( D_t \) and \( M_t \), we conclude that \( \{u_t\} \) are uniformly bounded.

**Corollary E.2.** Let \( v_t, e_t, u_t \) and \( D_t \) be the optimal solutions produced by Algorithm [7].

1. \( \hat{\ell}(z_t, D_t, v_t, e_t) \) and \( \ell(z_t, D_t) \) are both uniformly bounded.

2. \( \frac{1}{t} \hat{h}(Z, D, U) \) is uniformly bounded.

3. The surrogate function \( g_t(D_t) \) defined in Eq. (3.5) is uniformly bounded and Lipschitz.

**Proof.** To see the uniform boundedness of \( \hat{\ell}(z_t, D_t, v_t, e_t) \), we just need to examine the definition of \( \ell(z_t, D_t, v_t, e_t) \) (see Eq. (2.3)) and notice that \( z_t, D_t, v_t \) and \( e_t \) are all uniformly bounded. This implies that \( \ell(z_t, D_t, v_t, e_t) \) is uniformly bounded and so is \( \ell(z_t, D_t) \).

Similarly, we can show that \( \frac{1}{t} \hat{h}(Z, D, U) \) is uniformly bounded.

The uniform boundedness of \( g_t(D_t) \) follows immediately as \( \hat{\ell}(z_t, D_t, v_t, e_t) \) and \( \frac{1}{t} \hat{h}(Z, D, U) \) are both uniformly bounded. To show that \( g_t(D) \) is Lipschitz, we show that the gradient of \( g_t(D) \) is uniformly bounded for all \( D \in D \).

\[
\| \nabla g_t(D) \|_F = \left\| \lambda_1 D A_t + \frac{\lambda_3}{t} I - \lambda_1 \frac{B_t}{t} - \frac{\lambda_3}{t} M_t \right\|_F \\
\leq \lambda_1 \|D\|_F \left\| \frac{A_t}{t} \right\|_F + \left\| \frac{\lambda_3}{t} I \right\|_F + \lambda_1 \left\| \frac{B_t}{t} \right\|_F \\
+ \left\| \frac{\lambda_3}{t} M_t \right\|_F.
\]

Notice that each term on the right side of the inequality is uniformly bounded. Thus the gradient of \( g_t(D) \) is uniformly bounded and \( g_t(D) \) is Lipschitz.

**Proposition E.3.** Let \( D \in D \) and denote the minimizer of \( \hat{\ell}(z, D, v, e) \) as:

\[
\{v', e'\} = \arg \min_{v, e} \hat{\ell}(z, D, v, e).
\]

Then, the function \( \ell(z, L) \) is continuously differentiable and

\[
\nabla_D \ell(z, D) = (Dv' + e' - z)v'^T.
\]

Furthermore, \( \ell(z, \cdot) \) is uniformly Lipschitz.

**Proof.** By fixing the variable \( z \), the function \( \hat{\ell} \) can be seen as a mapping:

\[
\mathbb{R}^{d+p} \times D \to \mathbb{R}
\]

\[
([v; e], D) \mapsto \hat{\ell}(z, D, v, e)
\]

It is easy to show that for all \([v; e] \in \mathbb{R}^{d+p}, \hat{\ell}(z, \cdot, v, e) \) is differentiable. Also \( \ell(z, \cdot, \cdot) \) is continuous on \( \mathbb{R}^{d+p} \times D \). \( \nabla_D \ell(z, D, v, e) = (Dv + e - z)v'^T \) is continuous on \( \mathbb{R}^{d+p} \times D \). \( \forall D \in D \),

20
since $\ell(z, D, v, e)$ is strongly convex w.r.t. $v$, it has a unique minimizer $\{v', e'\}$. Thus Lemma [D.1] applies and we prove that $\ell(z, D)$ is differentiable in $D$ and
\[
\nabla_D \ell(z, D) = (Dv' + e' - z)v'^\top.
\]
Since every term in $\nabla_D \ell(z, D)$ is uniformly bounded (Assumption [I] and Proposition [4.2]), we conclude that the gradient of $\ell(z, D)$ is uniformly bounded, implying that $\ell(z, D)$ is uniformly Lipschitz w.r.t. $D$.

**Corollary E.4.** Let $f_i(D)$ be the empirical loss function defined in Eq. (2.6). Then $f_i(D)$ is uniformly bounded and Lipschitz.

**Proof.** Since $\ell(z, D)$ can be uniformly bounded (Corollary [E.2]), we only need to show that $\frac{1}{t} h(Z, D)$ is uniformly bounded. Note that we have derived the form $\frac{1}{t} h(Z, D)$ as follows:
\[
\frac{1}{t} h(Z, D) = \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \left\| D^\top \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_i \right)^{-1} z_i \right\|_2^2 + \frac{1}{2} \left\| \frac{\lambda_3}{2t^3} \left( \frac{1}{t} I + \frac{\lambda_3}{t} N_i \right)^{-1} \right\|_F^2
\]
where $N_i = \sum_{t=1}^{t} z_i z_i^\top$. Since every term in the above equation can be uniformly bounded, $h(Z, D)$ is uniformly bounded and so is $f_i(D)$.

To show that $f_i(D)$ is uniformly Lipschitz, we show that its gradient can be uniformly bounded.
\[
\nabla f_i(D) = \frac{1}{t} \sum_{i=1}^{t} \nabla \ell(z_i, D) + \frac{1}{t} \nabla h(Z, D)
\]
\[
= \frac{1}{t} \sum_{i=1}^{t} (D v_i + e_i - z_i) v_i^\top
\]
\[+ \frac{1}{t^2} \sum_{i=1}^{t} \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_i \right)^{-1} z_i z_i^\top \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_i \right)^{-1} D
\]
\[+ \frac{\lambda_3}{t^3} \left( \frac{1}{t} I + \frac{\lambda_3}{t} N_i \right)^{-2} D.
\]
Then the Frobenius norm of $\nabla f_i(D)$ can be bounded by:
\[
\| \nabla f_i(D) \|_F \leq \frac{1}{t} \sum_{i=1}^{t} \| D v_i + e_i - z_i \|_2 \cdot \| v_i \|_2
\]
\[+ \frac{1}{t^2} \sum_{i=1}^{t} \left\| \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_i \right)^{-1} \right\|_F^2 \cdot \| z_i \|_2 \cdot \| D \|_F
\]
\[+ \frac{\lambda_3}{t^3} \left\| \left( \frac{1}{t} I + \frac{\lambda_3}{t} N_i \right)^{-1} \right\|_F^2 \cdot \| D \|_F.
\]
One can easily check that the right side of the inequality is uniformly bounded. Thus $\| \nabla f_i(D) \|_F$ is uniformly bounded, implying that $f_i(D)$ is uniformly Lipschitz.

**E.2 Proof of P-Donsker**

**Proposition E.5.** Let $f'_i(D) = \frac{1}{t} \sum_{i=1}^{t} \ell(z_i, D)$ and $f(D)$ be the expected loss function defined in Eq. (2.10). Then we have
\[
\E[\sqrt{t} \| f'_i - f \|_\infty] = O(1).
\]
First, we bound the last four terms. We have in Proposition 4.2,

**Proof.** Here, to show the convergence of Theorem E.6 (Convergence of the surrogate function defined in Eq. (3.5)) is non-negative and the samples are drawn randomly. We define \( D \) (Proposition E.3). Thus, \( \{\ell(z, D)\} \) is P-Donsker (see the definition in Lemma D.2). Furthermore, as \( \ell(z, D) \) is non-negative and uniformly bounded, so is \( \ell^2(z, D) \). So we have \( \mathbb{E}_z[\ell^2(z, D)] \) being uniformly bounded. Since we have verified all the hypotheses in Lemma D.2 we obtain the result that

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

\[\square\]

**E.3 Proof of convergence of \( g_t(D) \)**

**Theorem E.6** (Convergence of the surrogate function \( g_t(D_t) \)). The surrogate function \( g_t(D_t) \) we defined in Eq. (3.5) converges almost surely, where \( D_t \) is the solution produced by Algorithm 7.

**Proof.** Note that \( g_t(D_t) \) can be viewed as a stochastic positive process since every term in \( g_t(D_t) \) is non-negative and the samples are drawn randomly. We define

\[
u_t = g_t(D_t).
\]

To show the convergence of \( u_t \), we need to bound the difference of \( u_{t+1} \) and \( u_t \):

\[
\begin{align*}
\nu_{t+1} - \nu_t &= g_{t+1}(D_{t+1}) - g_t(D_t) \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + g_{t+1}(D_t) - g_t(D_t) \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{1}{t+1} \ell(z_{t+1}, D_t) - \frac{1}{t+1} g'_t(D_t) \\
&\quad + \left\{ \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|_F^2 \\
&\quad - \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \right\} \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{f'_t(D_t) - g'_t(D_t)}{t+1} \\
&\quad + \frac{\ell(z_{t+1}, D_t) - f'_t(D_t)}{t+1} + \left\{ \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 \\
&\quad + \frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|_F^2 - \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 \\
&\quad - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \right\}.
\end{align*}
\]

(E.12)

Here,

\[
g'_t(D_t) = \frac{1}{t} \sum_{i=1}^{t} \ell(z_i, D, v_i, e_i).
\]

(E.13)

First, we bound the last four terms. We have

\[
\begin{align*}
\frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 - \frac{1}{t} \sum_{i=1}^{t} \|u_i\|^2 &\leq \frac{-1}{t(t+1)} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{1}{2(t+1)} \|u_{t+1}\|^2 \\
&\leq \frac{1}{2(t+1)} \|u_{t+1}\|^2.
\end{align*}
\]

(E.14)
And

\[
\frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|_F^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2
\]

\[
= -\frac{\lambda_3}{2t(t+1)} \|D_t - M_t\|_F^2 + \frac{\lambda_3}{2(t+1)} \|z_{t+1}u_{t+1}^\top\|_F^2
\]

\[
- \frac{\lambda_3}{t+1} \text{Tr}((D_t - M_t)^\top z_{t+1}u_{t+1}^\top)
\]

\[
= -\frac{\lambda_3}{2t(t+1)} \|D_t - M_t\|_F^2 + \frac{\lambda_3}{2(t+1)} \|z_{t+1}u_{t+1}^\top\|_F^2
\]

\[
- \frac{\lambda_3}{t+1} \left( \|z_{t+1}\|^2_2 + \frac{1}{\lambda_3} \|u_{t+1}\|^2_2 \right)
\]

\[
\leq \frac{1}{t+1} \left( \frac{\lambda_3}{2} \|z_{t+1}u_{t+1}^\top\|^2_F - (\lambda_3 \|z_{t+1}\|^2_2 + 1) \|u_{t+1}\|^2_2 \right)
\]

\[
\leq \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|^2_2 \|u_{t+1}\|^2_2 - \|u_{t+1}\|^2_2 \right)
\]

where the first equality is derived by the fact that $M_{t+1} = M_t + z_{t+1}u_{t+1}^\top$, and the second equality

Combining Eq. \((E.14)\) and Eq. \((E.15)\), we know that

\[
\frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2_2 - \frac{1}{t} \sum_{i=1}^{t} \|u_i\|^2_2
\]

\[
+ \frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|_F^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2
\]

\[
\leq \frac{1}{2(t+1)} \|u_{t+1}\|^2_2
\]

\[
+ \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|^2_2 \|u_{t+1}\|^2_2 - \|u_{t+1}\|^2_2 \right)
\]

\[
\leq \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|^2_2 \|u_{t+1}\|^2_2 - \frac{1}{2} \|u_{t+1}\|^2_2 \right)
\]

\[
\leq 0.
\]

Therefore,

\[
u_{t+1} - u_t
\]

\[
\leq g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{1}{t+1} \ell(z_{t+1}, D_t)
\]

\[
- \frac{1}{t+1} g_t'(D_t)
\]

\[
g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{f_t'(D_t) - g_t'(D_t)}{t+1}
\]

\[
+ \frac{\ell(z_{t+1}, D_t) - f_t'(D_t)}{t+1}
\]

\[
\leq \frac{\ell(z_{t+1}, D_t) - f_t'(D_t)}{t+1},
\]

where $f_t'(D)$ is defined in Proposition \((E.3)\) and the last inequality holds because $D_{t+1}$ is the mini-

mizer of $g_{t+1}(D)$ and $g_t'(D)$ is a surrogate function of $f_t'(D)$. 

23
Let $\mathcal{F}_t$ be the filtration of the past information. We take the expectation on the above equation conditioned on $\mathcal{F}_t$:

$$
\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t] \leq \frac{\mathbb{E}[\ell(z_{t+1}, D_t) \mid \mathcal{F}_t] - f'_t(D_t)}{t+1}
\leq \frac{f(D_t) - f'_t(D_t)}{t+1}
\leq \frac{\|f - f'_t\|_\infty}{t+1}.
$$

From Proposition E.5, we know

$$
\mathbb{E}[\|f - f'_t\|_\infty] = O\left(\frac{1}{\sqrt{t}}\right).
$$

Thus,

$$
\mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]] = \mathbb{E}[\max\{\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t], 0\}]
\leq \frac{c}{\sqrt{t}(t+1)},
$$

where $c$ is some constant.

Now let us define the index set

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

and the indicator

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

We have

$$
\sum_{t=1}^{\infty} \mathbb{E}[^{\delta_t}(u_{t+1} - u_t)] = \sum_{t \in \mathcal{T}} \mathbb{E}[u_{t+1} - u_t]
= \sum_{t \in \mathcal{T}} \mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]]
= \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]]^+
\leq +\infty.
$$

Thus, Lemma D.3 applies. That is, $g_t(D_t)$ is a quasi-martingale and converges almost surely. Moreover,

$$
\sum_{t=1}^{\infty} |\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]| < +\infty \text{ a.s.} \quad (E.16)
$$

E.4 Proof of Difference of $D_t$

**Proposition E.7.** Let $\{D_t\}_{t=1}^{\infty}$ be the basis sequence produced by the Algorithm[1] Then,

$$
\|D_{t+1} - D_t\|_F = O\left(\frac{1}{t}\right). \quad (E.17)
$$

**Proof.** According the strong convexity of $g_t(D)$ (Assumption3), we have,

$$
g_t(D_{t+1}) - g_t(D_t) \geq \frac{\beta_0}{2} \|D_{t+1} - D_t\|_F^2, \quad (E.18)
$$
On the other hand,
\[ g_t(D_{t+1}) - g_t(D_t) \]
\[ = g_t(D_{t+1}) - g_{t+1}(D_{t+1}) + g_{t+1}(D_{t+1}) - g_{t+1}(D_t) \]
\[ + g_{t+1}(D_t) - g_t(D_t) \]
\[ \leq g_t(D_{t+1}) - g_{t+1}(D_{t+1}) + g_{t+1}(D_t) - g_t(D_t) \]
\[ = G_t(D_{t+1}) - G_t(D_t). \] 

(E.19)

Note that the inequality is derived by the fact that \( g_{t+1}(D_{t+1}) - g_{t+1}(D_t) \leq 0 \), as \( D_{t+1} \) is the minimizer of \( g_{t+1}(D) \). And we denote \( g_t(D) - g_{t+1}(D) \) by \( G_t(D) \).

By a simple calculation, we obtain the gradient of \( G_t(D) \):
\[ \nabla G_t(D) \]
\[ = \nabla g_t(D) - \nabla g_{t+1}(D) \]
\[ = \frac{1}{t} [D(\lambda_1 A_t + \lambda_3 I) - (\lambda_1 B_t + \lambda_3 M_t)] \]
\[ - \frac{1}{t+1} [D(\lambda_1 A_{t+1} + \lambda_3 I) - (\lambda_1 B_{t+1} + \lambda_3 M_{t+1})] \]
\[ = \frac{1}{t} [D(\lambda_1 A_t + \lambda_3 I) - \frac{\lambda_1 t}{t+1} A_{t+1} - \lambda_3 \frac{t}{t+1} I) \]
\[ + \frac{\lambda_1 t}{t+1} B_{t+1} - \lambda_1 B_t + \frac{\lambda_3 t}{t+1} M_{t+1} - \lambda_3 M_t] \]
\[ = \frac{1}{t} [\nabla \lambda_1 (A_{t+1} - \lambda_1 v_{t+1} v_{t+1}^\top + \frac{\lambda_3}{t+1} I) \]
\[ + \lambda_1 (z_{t+1} - e_{t+1}) v_{t+1}^\top - \frac{\lambda_1}{t+1} B_{t+1} \]
\[ + \lambda_3 z_{t+1}^\top u_{t+1}^\top - \frac{\lambda_3}{t+1} M_{t+1} \]

So the Frobenius norm of \( \nabla G_t(D) \) follows immediately:
\[ \|\nabla G_t(D)\|_F \]
\[ \leq \frac{1}{t} \left[ \|D\|_F \left( \lambda_1 \left\| \frac{A_{t+1}}{t+1} \right\|_F + \lambda_1 \left\| v_{t+1} v_{t+1}^\top \right\|_F + \frac{\lambda_3}{t+1} \|I\|_F \right) \]
\[ + \lambda_1 \left\| (z_{t+1} - e_{t+1}) v_{t+1}^\top \right\|_F + \lambda_1 \left\| B_{t+1} \right\|_{t+1} \]
\[ + \lambda_3 \left\| z_{t+1}^\top u_{t+1}^\top \right\|_F + \frac{\lambda_3}{t+1} \|M_{t+1}\|_F \right] \]
\[ = \frac{1}{t} \left[ \|D\|_F \left( \lambda_1 \left\| \frac{A_{t+1}}{t+1} \right\|_F + \lambda_1 \left\| v_{t+1} v_{t+1}^\top \right\|_F \right) \]
\[ + \lambda_1 \left\| (z_{t+1} - e_{t+1}) v_{t+1}^\top \right\|_F + \lambda_1 \left\| B_{t+1} \right\|_{t+1} \]
\[ + \lambda_3 \left\| z_{t+1}^\top u_{t+1}^\top \right\|_F + \frac{\lambda_3}{t+1} \left[ \|I\|_F + \|M_{t+1}\|_F \right] \].

We know from Proposition (E.19) that all the terms in the above equation are uniformly bounded. Thus, there exist constants \( c_1, c_2 \) and \( c_3 \), such that
\[ \|\nabla G_t(D)\|_F \leq \frac{1}{t} [c_1 \|D\|_F + c_2] + \frac{c_3}{t(t+1)}. \]

According to the first order Taylor expansion,
\[ G_t(D_{t+1}) - G_t(D_t) \]
\[ = \text{Tr} \left( (D_{t+1} - D_t) \nabla G_t (\alpha D_t + (1 - \alpha) D_{t+1}) \right) \]
\[ \leq \|D_{t+1} - D_t\|_F \cdot \|\nabla G_t (\alpha D_t + (1 - \alpha) D_{t+1})\|_F, \]

25
where $\alpha$ is a constant between 0 and 1. According to Proposition 4.2, $D_t$ and $D_{t+1}$ are uniformly bounded, so $\alpha D_t + (1 - \alpha) D_{t+1}$ is uniformly bounded. Thus, there exists a constant $c_4$, such that

$$\|\nabla G_t (\alpha L_t + (1 - \alpha) L_{t+1})\|_F \leq \frac{c_4}{t} + \frac{c_3}{t(t+1)},$$

resulting that

$$G_t(D_{t+1}) - G_t(D_t) \leq \left( \frac{c_4}{t} + \frac{c_3}{t(t+1)} \right) \|D_{t+1} - D_t\|_F.$$

Combining Eq. (E.18), Eq. (E.19) and the above equation, we have

$$\|D_{t+1} - D_t\|_F = \frac{2c_4}{\beta_0} + \frac{1}{t} + \frac{2c_3}{\beta_0} \cdot \frac{1}{t(t+1)}.$$

\[ \square \]

E.5 Proof for convergence of $f_t(D_t)$

**Theorem E.8** (Convergence of $f_t(D_t)$). Let $f_t(D_t)$ be the empirical loss function defined in Eq. (2.6) and $D_t$ be the solution produced by the Algorithm 1. Let $b_t = g_t(D_t) - f_t(D_t)$. Then, $b_t$ converges almost surely to 0. Thus, $f_t(D_t)$ converges almost surely to the same limit of $g_t(D_t)$.

**Proof.** Let $f_t'(D)$ and $g_t'(D)$ be those defined in Proposition E.5 and Theorem 4.3 respectively. Then,

$$b_t = g_t(D_t) - f_t'(D_t)$$

$$= g_t(D_t) - f_t'(D_t) + \left( \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \right)$$

$$- \frac{1}{t^2} \sum_{i=1}^{t} \frac{1}{2} \left\|D_t^T \left( \frac{1}{\lambda_3 t} I + \frac{1}{t} N_t \right)^{-1} z_i \right\|^2_2$$

$$- \frac{\lambda_3}{2t^2} \left\| \left( \frac{1}{t} I + \frac{\lambda_3}{t} N_t \right)^{-1} D_t \right\|^2_F$$

$$= g_t'(D_t) - f_t'(D_t) + q_t(D_t),$$

where $g_t(D_t)$ denotes the last four terms. Combining Eq. (E.12) we have

$$\frac{b_t}{t+1} = \frac{g_t(D_t) - f_t'(D_t)}{t+1} + \frac{q_t(D_t)}{t+1}$$

$$= g_{t+1}(D_{t+1}) - g_t(D_t) + \frac{\ell(z_{t+1}, D_t) - f_t'(D_t)}{t+1}$$

$$+ u_t - u_{t+1} + \left\{ \frac{q_t(D_t)}{t+1} + \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 \right\}$$

$$+ \frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|_F^2 - \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2_2$$

$$- \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2.$$ 

Note that we naturally have

$$\frac{q_t(D_t)}{t+1} \leq \frac{1}{t(t+1)} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_3}{2t(t+1)} \|D_t - M_t\|_F^2$$

$$\leq \frac{1}{t(t+1)} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{c}{2t(t+1)},$$

26
where the second inequality holds as $D_t$ and $M_t$ are both uniformly bounded (see Proposition 4.2).

Also, from Eq. (E.14), we know

\[
\frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 - \frac{1}{t} \sum_{i=1}^{t} \|u_i\|^2 = \frac{-1}{t+1} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{1}{2(t+1)} \|u_{t+1}\|^2.
\]

And from Eq. (E.15)

\[
\frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|^2_F - \frac{\lambda_3}{2t} \|D_t - M_t\|^2_F \\
\leq \frac{1}{t+1} \left( \frac{\lambda_3}{2} \frac{\|z_{t+1}\|^2_2}{2} \frac{\|u_{t+1}\|^2_2}{2} - \frac{\|u_{t+1}\|^2_2}{2} \right)
\]

Thus,

\[
\frac{q_t(D_t)}{t+1} + \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_3}{2(t+1)} \|D_t - M_{t+1}\|^2_F \\
- \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|^2_F \\
\leq \frac{c}{2t(t+1)} + \frac{1}{2(t+1)} \|u_{t+1}\|^2_2 \\
+ \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \frac{\|z_{t+1}\|^2_2}{2} \frac{\|u_{t+1}\|^2_2}{2} - \frac{\|u_{t+1}\|^2_2}{2} \right) \\
= \frac{c}{2(t+1)} - \frac{1}{2(t+1)} \|u_{t+1}\|^2_2 - \frac{\lambda_3}{2(t+1)} \frac{\|z_{t+1}\|^2_2}{2} \frac{\|u_{t+1}\|^2_2}{2}
\]

Therefore,

\[
\frac{b_t}{t+1} \leq g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{\ell(z_{t+1}, D_t) - f_t'(D_t)}{t+1} \\
+ u_t - u_{t+1} + \frac{c}{2t(t+1)} \\
\leq \frac{\ell(z_{t+1}, D_t) - f_t'(D_t)}{t+1} + u_t - u_{t+1} + \frac{c}{2t(t+1)}.
\]

By taking the expectation conditioned on the past information $\mathcal{F}_t$, we have

\[
\frac{b_t}{t+1} \leq \frac{f(D_t) - f_t(D_t)}{t+1} + \mathbb{E}[u_t - u_{t+1} | \mathcal{F}_t] + \frac{c}{2t(t+1)} \\
\leq \frac{c_1}{\sqrt{t(t+1)}} + \mathbb{E}[u_t - u_{t+1} | \mathcal{F}_t] + \frac{c}{2t(t+1)},
\]

where the second inequality holds by applying Proposition E.5. Thus,

\[
\sum_{t=1}^{\infty} \frac{b_t}{t+1} \leq \sum_{t=1}^{\infty} \frac{c_1}{\sqrt{t(t+1)}} + \sum_{t=1}^{\infty} \mathbb{E}[u_t - u_{t+1} | \mathcal{F}_t] + \frac{c}{2t(t+1)} \\
\leq +\infty.
\]
Here, the last inequality is derived by applying Eq. (E.16).

Next, we examine the difference between $b_{t+1}$ and $b_t$:

$$|b_{t+1} - b_t| = |g_{t+1}(D_{t+1}) - f_{t+1}(D_{t+1}) - g_t(D_t) + f_t(D_t)|$$

$$\leq |g_{t+1}(D_{t+1}) - g_t(D_{t+1})| + |g_t(D_{t+1}) - g_t(D_t)|$$

$$+ |f_{t+1}(D_{t+1}) - f_t(D_{t+1})| + |f_t(D_{t+1}) - f_t(D_t)|.$$  \hspace{1cm} (E.20)

For the first term on the right side,

$$|g_{t+1}(D_{t+1}) - g_t(D_{t+1})| = |g_{t+1}'(D_{t+1}) - g_t'(D_{t+1}) + \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|^2$$

$$- \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2 + \frac{\lambda_3}{2(t+1)} \|D_{t+1} - M_t\|^2_F$$

$$- \frac{\lambda_3}{2t} \|D_{t+1} - M_t\|^2_F$$

$$\leq |g_{t+1}'(D_{t+1}) - g_t'(D_{t+1})| + \frac{1}{t(t+1)} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|^2$$

$$+ \frac{1}{2(t+1)} \|u_{t+1}\|^2 + \frac{\lambda_3}{2(t+1)} \|D_{t+1} - M_t\|^2_F$$

$$+ \frac{\lambda_3}{2(t+1)} \|z_{t+1}u_{t+1}^T\|^2_F$$

$$\leq \frac{c_1}{t+1}$$

$$\leq c_2 \frac{c_1}{t+1},$$

where $c_1$ and $c_2$ are some uniform constants. Note that $\zeta_1$ holds because all the $u_i$'s, $D_{t+1}$, $M_t$ and $z_{t+1}$ are uniformly bounded (see Proposition [4.2]), and $\zeta_2$ holds because $\ell(z_{t+1}, D_{t+1})$ and $g_t'(D_{t+1})$ are uniformly bounded (see Corollary [E.2]).

For the third term on the right side of Eq. (E.20), we can similarly derive

$$|f_{t+1}(D_{t+1}) - f_t(D_{t+1})|$$

$$\leq |f_{t+1}'(D_{t+1}) - f_t'(D_{t+1})| + \frac{c_3}{t+1}$$

$$= \left| \frac{1}{t+1} f((z_{t+1}, D_{t+1}) - f_{t+1}'(D_{t+1})) \right| + \frac{c_3}{t+1}$$

$$\leq \frac{c_3}{t+1},$$

where $c_3$ and $c_4$ are some uniform constants, and $\zeta_3$ holds as $\ell(z_{t+1}, D_{t+1})$ and $f_t'(D_{t+1})$ are both uniformly bounded (see Corollary [E.4]).
From Corollary E.2 and Corollary E.4, we know that both \( g_t(D) \) and \( f_t(D) \) are uniformly Lipschitz. That is, there exists uniform constants \( \kappa_1, \kappa_2 \), such that
\[
|g_t(D_{t+1}) - g_t(D_t)| \leq \kappa_1 \|D_{t+1} - D_t\|_F \leq \frac{\kappa_3}{t+1}
\]
\[
|f_t(D_{t+1}) - f_t(D_t)| \leq \kappa_2 \|D_{t+1} - D_t\|_F \leq \frac{\kappa_4}{t+1}.
\]

Here, \( \zeta_4 \) and \( \zeta_5 \) are derived by applying Proposition E.7 and \( \kappa_3 \) and \( \kappa_4 \) are some uniform constants. Finally, we have a bound for Eq. (E.20):
\[
|b_{t+1} - b_t| \leq \frac{\kappa_0}{t+1},
\]
where \( \kappa_0 \) is some uniform constant.

By applying Lemma D.4, we conclude that \( \{b_t\} \) converges to zero. That is,
\[
\lim_{t \to +\infty} g_t(D_t) - f_t(D_t) = 0.
\]

Since we have proved in Theorem 4.3 that \( g_t(D_t) \) converges almost surely, we conclude that \( f_t(D_t) \) converges almost surely to the same limit of \( g_t(D_t) \).

**Theorem E.9** (Convergence of \( f(D_t) \)). Let \( f(D) \) be the expected loss function we defined in Eq. (2.10) and let \( D_t \) be the optimal solution produced by Algorithm 7. Then \( f(D_t) \) converges almost surely to the same limit of \( f(D_t) \) (or \( g_t(D_t) \)).

**Proof.** According to the central limit theorem, we know that \( \sqrt{7}(f(D_t) - f(D_t)) \) is bounded, implying
\[
\lim_{t \to +\infty} f(D_t) - f(D_t) = 0, \quad a.s.
\]

**E.6 Proof of gradient of \( f(D) \)**

**Proposition E.10** (Gradient of \( f(D) \)). Let \( f(D) \) be the expected loss function which is defined in Eq. (2.10). Then, \( f(D) \) is continuously differentiable and \( \nabla f(D) = \mathbb{E}_u [\nabla_D \ell(z, D)] \). Moreover, \( \nabla f(D) \) is uniformly Lipschitz on \( D \).

**Proof.** We have shown in Proposition E.3 that \( \ell(z, D) \) is continuously differentiable, \( f(D) \) is also continuously differentiable and we have \( \nabla f(D) = \mathbb{E}_u [\nabla_D \ell(z, D)] \).

Next, we prove the Lipschitz of \( \nabla f(D) \). Let \( v'(z', D') \) and \( e'(z', D') \) be the minimizer of \( \hat{\ell}(z', D, v, e) \). Since \( \hat{\ell}(z, D, v, e) \) has a unique minimum and is continuous in \( z, D, v \) and \( e \), \( v'(z', D') \) and \( e'(z', D') \) is continuous in \( z \) and \( D \).

Let \( A = \{j \mid e_j' \neq 0\} \). According the first order optimality condition, we know that
\[
\frac{\partial \hat{\ell}(z, D, v, e)}{\partial e} = 0
\]
\[
\Rightarrow \lambda_1(z - Dv - e) \in \lambda_2 \|e\|_1
\]
\[
\Rightarrow |(z - Dv - e)| = \frac{\lambda_2}{\lambda_1}, \forall j \in A.
\]

Since \( z - Dv - e \) is continuous in \( z \) and \( D \), there exists an open neighborhood \( V \), such that for all \( (z'', D'') \in V \), if \( j \notin A \), then \( |(z'' - D''v'' - e'')| < \frac{\lambda_2}{\lambda_1} \) and \( e'' = 0 \). That is, the support set of \( e' \) will not change.

Let us denote \( H = [D I], r = [v^\top e^\top]^\top \) and define the function
\[
\hat{\ell}(z, H, r_A) = \frac{\lambda_1}{2} \|z - HA r_A\|_2^2 + \frac{1}{2} \|[I 0] r_A\|_2^2 + \lambda_2 \|[0 I] r_A\|_1
\]
Since \( \tilde{\ell}(z, D_A, \cdot) \) is strongly convex, there exists a uniform constant \( \kappa_1 \), such that for all \( r' \),

\[
\tilde{\ell}(z', H'_A, r'_A) - \tilde{\ell}(z', H'_A, r'_A) \\
\geq \kappa_1 \| r''_A - r'_A \|^2 \\
= \kappa_1 \left( \| v'' - v' \|^2 + \| e''_A - e'_A \|^2 \right).
\]

(E.21)

On the other hand,

\[
\tilde{\ell}(z', H'_A, r'_A) - \tilde{\ell}(z', H'_A, r'_A) \\
= \tilde{\ell}(z', H'_A, r'_A) - \tilde{\ell}(z'', H''_A, r''_A) + \tilde{\ell}(z'', H''_A, r''_A) \\
- \tilde{\ell}(z', H'_A, r'_A) \\
\leq \tilde{\ell}(z', H'_A, r'_A) - \tilde{\ell}(z'', H''_A, r''_A) + \tilde{\ell}(z'', H''_A, r''_A) \\
- \tilde{\ell}(z', H'_A, r'_A),
\]

(E.22)

where the last inequality holds because \( r'' \) is the minimizer of \( \tilde{\ell}(z', H'', r) \).

We shall prove that \( \tilde{\ell}(z', H'_A, r_A) - \tilde{\ell}(z'', H''_A, r_A) \) is Lipschitz w.r.t. \( r \), which implies the Lipschitz of \( v'(z, D) \) and \( e'(z, D) \).

\[
\nabla_r \left( \tilde{\ell}(z', H'_A, r_A) - \tilde{\ell}(z'', H''_A, r_A) \right) \\
= \lambda_1 \left[ H''_A (H'_A - H'_A) + (H'_A - H'_A)^\top H''_A + H''_A (z'' - z') \\
+ (H''_A - H''_A)^\top z' \right].
\]

Note that \( \| H'_A \|_F, \| H''_A \|_F \) and \( z'' \) are all uniformly bounded. Hence, there exists uniform constants \( c_1 \) and \( c_2 \), such that

\[
\| \nabla_r \left( \tilde{\ell}(z', H'_A, r_A) - \tilde{\ell}(z'', H''_A, r_A) \right) \|_2 \leq c_1 \| H'_A - H''_A \|_F \\
+ c_2 \| z' - z'' \|_2,
\]

which implies that \( \tilde{\ell}(z', H'_A, r_A) - \tilde{\ell}(z'', H''_A, r_A) \) is Lipschitz with Lipschitz constant \( c(H'_A, H''_A, z', z'') = c_1 \| H'_A - H''_A \|_F + c_2 \| z' - z'' \|_2 \). Combining this fact with Eq. (E.21) and Eq. (E.22), we obtain

\[
\kappa_1 \| r''_A - r'_A \|^2 \\
\leq c(H'_A, H''_A, z', z'') \| r''_A - r'_A \|_2.
\]

Therefore, \( r(z, D) \) is Lipschitz and so are \( v(z, D) \) and \( e(z, D) \). Note that according to Proposition 3.3,

\[
\nabla f(D') - \nabla f(D'') \\
= \mathbb{E}_z \left[ (H'z' - z)v'^\top - (H''z' - z)v'^\top \right] \\
= \mathbb{E}_z \left[ H'z'(v' - v'') + (H' - H'')v'v'^\top + H''(r' - r'')v'^\top + z(v'' - v') \right].
\]

So,

\[
\| \nabla f(D') - \nabla f(D'') \|_F \\
\leq \mathbb{E}_z \left[ \| H'z' \|_2 \| v' - v'' \|_2 + \| H' - H'' \|_F \| r'v'^\top \|_F \\
+ \| H'' \|_F \| r' - r'' \|_2 \| v'' \|_2 + \| z \|_2 \| v' - v'' \|_2 \right] \\
\leq c_2 \mathbb{E}_z \left[ \gamma_1 + \gamma_2 \| z \|_2 \right] \| H' - H'' \|_F \]

\[
\leq \gamma_0 \| D' - D'' \|_F.
\]

30
where $\gamma_0$, $\gamma_1$ and $\gamma_2$ are all uniform constants. Here, $\zeta_1$ holds because for any function $s(z)$, we have $\|E_x[s(z)]\|_F \leq E_x[\|s(z)\|_F]$. $\zeta_2$ is derived by using the result that $r(z, H)$ and $v(z, H)$ are both Lipschitz and $H'$, $H''$, $r'$, $r''$, $v'$ and $v''$ are all uniformly bounded. $\zeta_3$ holds because $\mathbf{z}$ is uniformly bounded and actually $\|H' - H''\|_F = \|D' - D''\|_F$. Thus, we complete the proof.

E.7 Proof of stationary point

**Theorem E.11** (Convergence of $D_t$). Let $\{D_t\}$ be the optimal basis produced by Algorithm[7] and let $f(D)$ be the expected loss function defined in Eq. (2.10). Then $D_t$ converges to a stationary point of $f(D)$ when $t$ goes to infinity.

**Proof.** Since $\frac{1}{h}A_t$ and $\frac{1}{h}B_t$ are uniformly bounded (Proposition 4.2), there exist sub-sequences of $\{\frac{1}{t}A_t\}$ and $\{\frac{1}{t}B_t\}$ that converge to $A_\infty$ and $B_\infty$ respectively. Then $D_t$ will converge to $D_\infty$. Let $W$ be an arbitrary matrix in $\mathbb{R}^{p \times d}$ and $\{h_k\}$ be any positive sequence that converges to zero. As $g_t$ is a surrogate function of $f_t$, for all $t$ and $k$, we have

$$g_t(D_t + h_kW) \geq f_t(D_t + h_kW).$$

Let $t$ tend to infinity, and note that $f(D) = \lim_{t \to \infty} f_t(D)$, we have

$$g_\infty(D_\infty + h_kW) \geq f(D_\infty + h_kW).$$

Note that the Lipschitz of $\nabla f$ indicates that the second derivative of $f(D)$ is uniformly bounded. By a simple calculation, we can also show that it also holds for $g_t(D)$. This fact implies that we can take the first order Taylor expansion for both $g_t(D)$ and $f(D)$ even when $t$ tends to infinity (because the second order derivatives of them always exist). That is,

$$\text{Tr}(h_kW^\top \nabla g_\infty(D_\infty)) + o(h_kW) \geq \text{Tr}(h_kW^\top \nabla f(D_\infty)) + o(h_kW)$$

By multiplying $\frac{1}{h_k\|W\|_F}$ on both sides and note that $\{h_k\}$ is a positive sequence, it follows that

$$\text{Tr}(\frac{1}{\|W\|_F} W^\top \nabla g_\infty(D_\infty)) + \frac{o(h_kW)}{h_k\|W\|_F} \geq \text{Tr}(\frac{1}{\|W\|_F} W^\top \nabla f(D_\infty)) + \frac{o(h_kW)}{h_k\|W\|_F}$$

Now let $k$ go to infinity,

$$\text{Tr}(\frac{1}{\|W\|_F} W^\top \nabla g_\infty(D_\infty)) \geq \text{Tr}(\frac{1}{\|W\|_F} W^\top \nabla f(D_\infty))$$

Note that this inequality holds for any matrix $W \in \mathbb{R}^{p \times d}$, so we actually have

$$\nabla g_\infty(D_\infty) = \nabla f(D_\infty).$$

As $D_\infty$ is the minimizer of $g_\infty(D)$, we have

$$\nabla f(D_\infty) = \nabla g_\infty(D_\infty) = 0.$$
References

[1] Haim Avron, Satyen Kale, Vikas Sindhwani, and Shiva P Kasiviswanathan. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In Proceedings of the 29th International Conference on Machine Learning (ICML-12), pages 1231–1238, 2012.

[2] Francis R Bach. Consistency of trace norm minimization. The Journal of Machine Learning Research, 9:1019–1048, 2008.

[3] Dimitri P Bertsekas. Nonlinear programming. 1999.

[4] J Frédéric Bonnans and Alexander Shapiro. Optimization problems with perturbations: A guided tour. SIAM review, 40(2):228–264, 1998.

[5] Léon Bottou. Online learning and stochastic approximations. On-line learning in neural networks, 17(9), 1998.

[6] Samuel Burer and Renato DC Monteiro. Local minima and convergence in low-rank semidefinite programming. Mathematical Programming, 103(3):427–444, 2005.

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

[8] Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? Journal of the ACM (JACM), 58(3):11, 2011.

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

[10] Emmanuel J Candès and Terence Tao. Decoding by linear programming. Information Theory, IEEE Transactions on, 51(12):4203–4215, 2005.

[11] Venkat Chandrasekaran, Sujay Sanghavi, Pablo A Parrilo, and Alan S Willsky. Rank-sparsity incoherence for matrix decomposition. SIAM Journal on Optimization, 21(2):572–596, 2011.

[12] Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2790–2797. IEEE, 2009.

[13] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. A rank minimization heuristic with application to minimum order system approximation. In American Control Conference, 2001. Proceedings of the 2001, volume 6, pages 4734–4739. IEEE, 2001.

[14] Jiashi Feng, Huan Xu, and Shuicheng Yan. Online robust pca via stochastic optimization. In NIPS, pages 404–412, 2013.

[15] Elaine T Hale, Wotao Yin, and Yin Zhang. Fixed-point continuation for $\ell_1$-minimization: Methodology and convergence. SIAM Journal on Optimization, 19(3):1107–1130, 2008.

[16] Elad Hazan. Sparse approximate solutions to semidefinite programs. In LATIN 2008: Theoretical Informatics, pages 306–316. Springer, 2008.

[17] Cho-Jui Hsieh and Peder Olsen. Nuclear norm minimization via active subspace selection. In Proceedings of The 31st International Conference on Machine Learning, pages 575–583, 2014.

[18] Daniel Hsu, Sham M Kakade, and Tong Zhang. Robust matrix decomposition with sparse corruptions. Information Theory, IEEE Transactions on, 57(11):7221–7234, 2011.

[19] Martin Jaggi, Marek Sulovsk, et al. A simple algorithm for nuclear norm regularized problems. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 471–478, 2010.

[20] Prateek Jain, Praneeth Netrapalli, and Sujay Sanghavi. Low-rank matrix completion using alternating minimization. In Proceedings of the 45th annual ACM symposium on Symposium on theory of computing, pages 665–674. ACM, 2013.

[21] Zhouchen Lin, Minming Chen, and Yi Ma. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. arXiv preprint arXiv:1009.5055, 2010.

[22] Guangcan Liu and Ping Li. Advancing matrix completion by modeling extra structures beyond low-rankness. arXiv preprint arXiv:1404.4646, 2014.
[23] Guangcan Liu, Zhouchen Lin, Shuicheng Yan, Ju Sun, Yong Yu, and Yi Ma. Robust recovery of subspace structures by low-rank representation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(1):171–184, 2013.

[24] Guangcan Liu, Zhouchen Lin, and Yong Yu. Robust subspace segmentation by low-rank representation. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 663–670, 2010.

[25] Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro. Online learning for matrix factorization and sparse coding. *The Journal of Machine Learning Research*, 11:19–60, 2010.

[26] Praneeth Netrapalli, Niranjan U. N, Sujay Sanghavi, Animashree Anandkumar, and Prateek Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems 27*, pages 1107–1115, 2014.

[27] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. pages 849–856, 2001.

[28] Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

[29] Ohad Shamir and Shai Shalev-Shwartz. Matrix completion with the trace norm: learning, bounding, and transducing. *Journal of Machine Learning Research*, 15(1):3401–3423, 2014.

[30] Mahdi Soltanolkotabi, Emmanuel J Candès, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

[31] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[32] Huan Xu, Constantine Caramanis, and Shie Mannor. Principal component analysis with contaminated data: The high dimensional case. In *COLT*, pages 490–502, 2010.