A. Overview of the Appendix

In Section B, we derive the optimal solution $U^*$ for (2.4), which was used in Section 2. Next, we clarify the implementation details of Algorithm 1 in Section C. We illustrate more experiments in Section D. In Section F, we prove our theoretical results and some preliminary lemmas are given in Section E.

B. Optimal solution $U^*$ for (2.4)

The optimal solution $U$ for (2.4) is given by the first order optimality condition:

$$\frac{\partial h(Y, D, U)}{\partial U} = U + \lambda_3(Y^TYU - Y^TD) = 0,$$

which implies

$$U^* = \left(\frac{1}{\lambda_3} I_p + Y^TY\right)^{-1} Y^TD = \lambda_3 \left(\sum_{j=0}^{+\infty} (-\lambda_3 Y^TY)^j Y^TD\right) = \lambda_3 Y^T \left[\sum_{j=0}^{+\infty} (-\lambda_3 Y^TY)^j\right] D = \lambda_3 Y^T \left(\frac{1}{\lambda_3} I_p + YY^T\right)^{-1} D.$$

Then, its transpose is given by

$$U^{*\top} = \left(\frac{1}{\lambda_3} I_p + \sum_{i=1}^{n} y_i y_i^\top\right)^{-1} Y.$$

Note that $u_i$ is the column of $U^\top$. So for each $i \in [n],$$u_i^* = \left(\frac{1}{\lambda_3} I_p + \sum_{i=1}^{n} y_i y_i^\top\right)^{-1} y_i = \frac{1}{n} \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n N_n}\right)^{-1} y_i,

where we denote $N_n = \sum_{i=1}^{n} y_i y_i^\top$.

Also, we have

$$YU^*\top = Y \left(\frac{1}{\lambda_3} I_p + Y^TY\right)^{-1} Y^TD = \lambda_3(1 + \lambda_3 Y^TY)^{-1} Y^TD = \lambda_3 Y \left[\sum_{j=0}^{+\infty} (-\lambda_3 Y^TY)^j\right] Y^TD = \lambda_3 \sum_{j=0}^{+\infty} (-\lambda_3)^j (Y^TY)^{j+1} D = D - (I_p + \lambda_3 YY^\top)^{-1} D = D - \left(\frac{1}{n} I_p + \frac{\lambda_3}{n N_n}\right)^{-1} D.$$

Thus,

$$h(Y, D) = \frac{1}{2} \left\|\frac{1}{n} D^\top \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n N_n}\right) - \frac{1}{n} D^\top \right\|_F^2 + \frac{\lambda_3}{2} \left\|\frac{1}{n} I_p + \frac{\lambda_3}{n N_n}\right\|_F^2 = \frac{1}{n^2} \sum_{i=1}^{n} \left\|D^\top \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n N_n}\right) - \frac{1}{n} D^\top \right\|_F^2 + \frac{\lambda_3}{2n^2} \left\|\frac{1}{n} I_p + \frac{\lambda_3}{n N_n}\right\|_F^2.$$

C. Algorithm Details

**Algorithm 2** Solving $v$ and $e$

**Require**: $D \in \mathbb{R}^{p \times d}$, $z \in \mathbb{R}^p$, parameters $\lambda_1$ and $\lambda_2$

**Ensure**: 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 $e$:

$$e = S_{\lambda_2/\lambda_1}[z - Dv].$$

5: **until** convergence

For Algorithm 2, we set a threshold $\epsilon = 10^{-3}$. Let $\{v', e'\}$ and $\{v'', e''\}$ be the two consecutive iterates. If the maximum of $\|v' - v''\|_2 / \|v\|_2$ and $\|e' - e''\|_2 / \|e\|_2$ is less than $\epsilon$, then we stop Algorithm 2.
online Low-Rank Subspace Clustering by Basis Dictionary Pursuit

Algorithm 3 Solving $D$

Require: $D \in \mathbb{R}^{p \times d}$ in the previous iteration, accumulation matrix $M$, $A$ and $B$, parameters $\lambda_1$ and $\lambda_3$.
Ensure: Optimal $D$ (updated).
1: Denote $A = \lambda_1 A + \lambda_3 I$ and $\hat{B} = \lambda_1 B + \lambda_3 M$.
2: repeat
3: for $j = 1$ to $d$ do
4: Update the $j$th column of $D$:
\[
d_j \leftarrow d_j - \frac{1}{A_{jj}} \left( D \hat{a}_j - \hat{b}_j \right)
\]
5: end for
6: until convergence

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. This is also observed in Mairal et al. (2010).

D. More Experiments

We also investigate the performance of subspace clustering on MNIST-7K and MNIST-10K. In this way, one can see how the computational time changes with the number of samples.

Table 4. Clustering accuracy (%) and computational time (seconds).

<table>
<thead>
<tr>
<th></th>
<th>OLRC</th>
<th>ORPCA</th>
<th>LRR</th>
<th>LRR2</th>
<th>SSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mushrooms</td>
<td>85.09</td>
<td>65.26</td>
<td>58.44</td>
<td>56.38</td>
<td>54.16</td>
</tr>
<tr>
<td>DNA</td>
<td>67.11</td>
<td>53.11</td>
<td>44.01</td>
<td>45.32</td>
<td>52.23</td>
</tr>
<tr>
<td>Protein</td>
<td>43.30</td>
<td>40.22</td>
<td>40.31</td>
<td>40.00</td>
<td>44.27</td>
</tr>
<tr>
<td>USPS</td>
<td>65.95</td>
<td>55.70</td>
<td>52.98</td>
<td>58.69</td>
<td>47.58</td>
</tr>
<tr>
<td>MNIST-7K</td>
<td>58.04</td>
<td>55.40</td>
<td>54.77</td>
<td>54.27</td>
<td>45.56</td>
</tr>
<tr>
<td>MNIST-10K</td>
<td>56.79</td>
<td>54.66</td>
<td>55.15</td>
<td>53.67</td>
<td>44.90</td>
</tr>
<tr>
<td>MNIST-20K</td>
<td>57.74</td>
<td>54.10</td>
<td>55.23</td>
<td>54.53</td>
<td>43.91</td>
</tr>
</tbody>
</table>

E. Proof Preliminaries

Lemma 3 (Corollary of Thm. 4.1 (Bonnans & Shapiro, 1998)). 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 4 (Corollary of Donsker theorem (van der 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 $\Theta$ of $\mathbb{R}^d$. Suppose that there exists a constant $K$ such that
\[
|f_{\theta_1}(x) - f_{\theta_2}(x)| \leq K \| \theta_1 - \theta_2 \|_2^2,
\]
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 $P_n f, \mathbb{P} f$ and $\mathbb{G}_n f$ as
\[
P_n f = \frac{1}{n} \sum_{i=1}^n f(X_i), \quad \mathbb{P} f = \mathbb{E}[f(X)], \quad \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n f - \mathbb{P} f).
\]
Let us also suppose that for all $f$, $\mathbb{P} f^2 < 2^2$ and $\| f \|_\infty < M$ and that the random elements $X_1, X_2, \cdots$ are Borel measurable. Then, we have
\[
\mathbb{E} \| \mathbb{G} \|_F = O(1),
\]
where $\| \mathbb{G} \|_F = \sup_{f \in F} |\mathbb{G}_n f|$.

Lemma 5 (Sufficient condition of convergence for a stochastic process (Bottou, 1998)). Let $(\Omega, \mathcal{F}, \mathbb{P})$ 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
\[
\delta_t = \begin{cases} 1 & \text{if } \mathbb{E}[u_{t+1} - u_t | \mathcal{F}_t] > 0, \\ 0 & \text{otherwise.} \end{cases}
\]
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,
\[
\sum_{t=1}^\infty |\mathbb{E}[u_{t+1} - u_t | \mathcal{F}_t]| < +\infty \text{ a.s.}
\]

Lemma 6 (Lemma 8 from Mairal et al. (2010)). 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, $\lim_{t \to +\infty} b_t = 0$.

F. Proof Details

F.1. Proof of Boundedness

Proposition 7. Let $\{ u_t \}$, $\{ v_t \}$, $\{ e_t \}$ and $\{ D_t \}$ be the optimal solutions produced by Algorithm 1. 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 of solving $v$ and $e$. As the trivial solution $\{v_t', e_t'\} = \{0, z_t\}$ are feasible, we have

$$\bar{t}_1(z_t, D_{t-1}, v_t', e_t') = \lambda_2 \|z_t\|_1.$$  

Therefore, the optimal solution should satisfy:

$$\frac{\lambda_1}{2} \|z_t - D_{t-1} v_t - e_t\|^2_t + \frac{1}{2} \|v_t\|^2_t + \lambda_2 \|e_t\|_1 \leq \lambda_2 \|z_t\|_1,$$

which implies

$$\frac{1}{2} \|v_t\|^2_t \leq \lambda_2 \|z_t\|_1,$$

$$\lambda_2 \|e_t\|_1 \leq \lambda_2 \|z_t\|_1.$$

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 each $v_i$, $e_i$, 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. $M_t = \sum_{i=1}^t y_i u_i^\top$ (definition of $M_t$).

2. $u_t = (\|y_t\|^2_t + \frac{1}{\lambda_3})^{-1}(D_{t-1} - M_{t-1})^\top y_t$ (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 for $D_t$).

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 (3.4), we know that

$$M_{t+1} - M_t = y_{t+1} u_{t+1}^\top$$

$$= \left(\|y_{t+1}\|^2 + \frac{1}{\lambda_3}\right)^{-1} y_{t+1} y_{t+1}^\top (D_t - M_t)$$

$$= P_t D_t - P_t M_t$$

$$= P_t(\lambda_1 B_t + \lambda_3 M_t) Q_t^{-1} - P_t M_t,$$

where

$$P_t = \left(\|y_{t+1}\|^2 + \frac{1}{\lambda_3}\right)^{-1} y_{t+1} y_{t+1}^\top.$$

By multiplying $Q_t$ on both sides of (F.2), we have

$$M_{t+1} = (M_t - \lambda_1 P_t M_t A_t Q_t^{-1}) + \lambda_3 P_t B_t Q_t^{-1}.$$  

(F.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,

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

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

$$= -\frac{1}{\lambda_1} \left[ I_d + \left( -\frac{\lambda_1}{\lambda_3} A_t \right)^{-1} \right]^{-1} + \frac{1}{\lambda_1} I_d.$$

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

$$M_{t+1} = (I_d - P_t) M_t$$

$$+ P_t M_t \left( I_d + \left( -\frac{\lambda_1}{\lambda_3} A_t \right)^{-1} \right) + \lambda_3 P_t B_t Q_t^{-1}.$$

(F.4)
Online Low-Rank Subspace Clustering by Basis Dictionary Pursuit

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

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

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,

$$
\left\|\left( \frac{1}{t} Q_t \right)^{-1} \right\| = \frac{1}{\sigma_{\min} \left( \frac{1}{t} Q_t \right)}
= \frac{1}{\sigma_{\min} \left( \frac{1}{t} A_t + \frac{1}{t} I_d \right)}
= \frac{\lambda \alpha + \lambda_1 \sigma_{\min} \left( \frac{1}{t} A_t \right)}{1}
\leq \frac{1}{\lambda_3 + \lambda_1 \beta_0}.
$$

Thus, $\lambda_1 P_t B_t Q_t^{-1}$ is uniformly bounded by a constant, say $c_2$. That is,

$$
\left\|\lambda_1 P_t B_t Q_t^{-1}\right\| \leq c_2. \tag{F.5}
$$

It follows that $W_t$ can be bounded:

$$
\|W_t\| \leq \|P_t\| \cdot \|M_t\| \cdot \left\|(I_d + \frac{\lambda_1}{\lambda_3} A_t)\right\|^{-1} + c_2
\leq \frac{\alpha_2 \lambda_3}{\alpha_2^2 + 1} \cdot \frac{\lambda_3}{\lambda_3 + \lambda_1 \beta_0} \|M_t\| + c_2 \tag{F.6}
$$

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 $\frac{1}{\tau} \sum_{i=1}^{\tau} z_i z_i^\top$ should uniformly span the diagonal. Thus, there exists a constant $\tau$, such that for all $i > 0$, $\frac{1}{\tau} \sum_{i=1}^{\tau} z_i z_i^\top$ is uniformly bounded away from zero with high probability.

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

$$
\frac{c_3 \tau}{t^*} (\frac{1}{\alpha_0} + 1) \leq 0.5. \tag{F.7}
$$

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

$$
\|W_t\| \leq w^*, \tag{F.8}
$$

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 (F.4) that for all $t > 0$,

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

Thus,

$$
\|M_{t+1}\|
= \left\|\prod_{i=1}^{t} (I_p - P_i) M_1 + \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} (I_p - P_i) W_j + W_t \right\|
\leq \left\|\prod_{i=1}^{t} (I_p - P_i) M_1 \right\| + \sum_{j=1}^{t-1} \prod_{i=j+1}^{t} (I_p - P_i) W_j + W_t
\leq \frac{\tau}{t} \left\|\prod_{i=1}^{t} (I_p - P_i) \right\| \cdot \|M_1\| + \tau w^*
\leq (1 - \alpha_0) m_1 + \tau w^*.
$$

Here, $\zeta_1$ holds because $\left\|\prod_{i=j+1}^{t} (I_p - P_i) \right\| \leq 1$ for all $j \in [\tau - 1]$. $\zeta_2$ holds because the singular values of $P_i$’s have span over the diagonal so the largest singular value of $\prod_{i=j+1}^{\tau} (I_p - P_i)$ is $1 - \alpha_0$, where $\alpha_0$ is the lower bound for all $z_t$'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 be 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 $M_t$’s, we can bound them as follows:
\[
\|M_t\| \leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*, \quad \forall t = 1, 2, \cdots, t^*. \quad (F.9)
\]
From (F.8) and (F.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,
\[
\|M_{t^*+1}\| \leq \|M_{k^*(\tau + 1)}\| + (t^* + 1 - k^*(\tau + 1))w^*
\]
\[
\leq m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*. \quad (F.10)
\]
For $\|W_{t^*+1}\|$, from (F.6), we know
\[
\|W_{t^*+1}\| \leq \frac{c_3}{t^* + 1} \|M_{t^*+1}\| + c_2
\]
\[
\leq \frac{c_3}{t^* + 1} (m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*) + c_2
\]
\[
= \frac{c_3 m_1}{t^* + 1} + \frac{c_3 \tau}{t^* + 1} (\frac{1}{\alpha_0} + 1) w^* + c_2
\]
\[
\leq \frac{c_3 m_1}{t^* + 1} + 0.5 w^* + c_2
\]
\[
\leq w^*. \quad (F.11)
\]
Here, $\zeta_1$ is derived by utilizing (F.7) and $\zeta_2$ is derived by (F.8).

From (F.10) and (F.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 (F.1), we know that
\[
D_t = \lambda_1 B_t (\lambda_1 A_t + \lambda_3 I_d)^{-1} + \lambda_3 M_t (\lambda_1 A_t + \lambda_3 I_d)^{-1}
\]
\[
= \lambda_1 \left( \frac{1}{t} B_t \right) \left( \frac{\lambda_1}{t} A_t + \frac{\lambda_3}{t} I_d \right)^{-1}
\]
\[
+ \lambda_3 M_t \left( \frac{\lambda_1}{t} A_t + \frac{\lambda_3}{t} I_d \right)^{-1}.
\]
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 8.** Let $v_t$, $e_t$, $u_t$ and $D_t$ be the optimal solutions produced by Algorithm 1.

1. $\tilde{\ell}(z_t, D_t, v_t, e_t)$ and $\ell(z_t, D_t)$ are uniformly bounded.
2. $\frac{1}{t} \tilde{h}(Z, D, U)$ is uniformly bounded.
3. The surrogate function $g_t(D_t)$ defined in (3.5) is uniformly bounded and Lipschitz.

**Proof.** To show Claim 1, we just need to examine the definition of $\tilde{\ell}(z_t, D_t, v_t, e_t)$ 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 $\ell(z_t, D_t)$. Likewise, we show that $\frac{1}{t} \tilde{h}(Z, D, U)$ is uniformly bounded.

The uniform boundedness of $g_t(D_t)$ follows immediately as $\ell(z_t, D_t, v_t, e_t)$ and $\frac{1}{t} \tilde{h}(Z, D, U)$ 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 $D \in D$.

\[
\|\nabla g_t(D)\|_F \leq \||\lambda_1 D \left( \frac{A_t}{t} + \frac{\lambda_3}{t} I_d \right) - \lambda_1 B_t - \frac{\lambda_3}{t} M_t\|_F
\]
\[
\leq \lambda_1 \|D\|_F \left( \left\| \frac{A_t}{t} \right\|_F + \left\| \frac{\lambda_3}{t} I_d \right\|_F \right)
\]
\[
+ \lambda_1 \left\| \frac{B_t}{t} \right\|_F + \left\| \frac{\lambda_3}{t} M_t \right\|_F.
\]
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 9. Let $D \in \mathcal{D}$ and denote the minimizer of $\ell(z, D, v, e)$ as:

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

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 $\ell$ can be seen as a mapping:

$$\mathbb{R}^{d+p} \times \mathcal{D} \to \mathbb{R},$$ 

$$(v, e, D) \mapsto \ell(z, D, v, e).$$

It is easy to show that for all $[v, e] \in \mathbb{R}^{d+p}, \ell(z, \cdot, v, e)$ is differentiable. Also $\ell(z, \cdot, v, e)$ is continuous on $\mathbb{R}^{d+p} \times \mathcal{D}$. $\nabla_D \ell(z, D, v, e) = (Dv + e - z)v'^T$ is continuous on $\mathbb{R}^{d+p} \times \mathcal{D}$. Since $\ell(z, D, v, e)$ is strongly convex w.r.t. $v$, it has a unique minimizer $\{v', e'\}$. Thus Lemma 3 applies and we prove that $\ell(z, D)$ is differentiable in $D$ and

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

Since every term in $\nabla_D \ell(z, D)$ is uniformly bounded (Assumption 1 and Proposition 7), 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 10. Let $f_i(D)$ be the empirical loss function defined in (2.6). Then $f_i(D)$ is uniformly bounded and Lipschitz.

**Proof.** Since $\ell(z, L)$ can be uniformly bounded (Corollary 8), we only need to show that $\frac{1}{t}h(Z, D)$ is uniformly bounded. Note that we have derived the form for $h(Z, D)$ as follows:

$$\frac{1}{t}h(Z, D) = \frac{1}{t^3} \sum_{i=1}^{t} \left[ D^T \left( \frac{1}{\lambda I_p} + \frac{1}{t} N_i \right)^{-1} z_i \right]_2^2$$

$$+ \frac{\lambda_3}{2t^3} \left[ \left( \frac{1}{t} I_p + \frac{\lambda_3}{t} N_i \right)^{-1} D \right]_F^2,$$

where $N_i = \sum_{i=1}^{t} z_i z_i^T$. 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:

$$\nabla f_i(D)$$

$$= \frac{1}{t} \sum_{i=1}^{t} \nabla \ell(z_i, D) + \frac{1}{t} \nabla h(Z, D)$$

$$= \frac{1}{t} \sum_{i=1}^{t} (Dv_i + e_i - z_i)v_i^T$$

$$+ \frac{1}{t^3} \sum_{i=1}^{t} \left( \frac{1}{\lambda I_p} + \frac{1}{t} N_i \right)^{-1} z_i z_i^T \left( \frac{1}{\lambda I_p} + \frac{1}{t} N_i \right)^{-1} D$$

$$+ \frac{\lambda_3}{t^3} \left( \frac{1}{t} I_p + \frac{\lambda_3}{t} N_i \right)^{-2} D.$$

Then the Frobenius norm of $\nabla f_i(D)$ can be bounded by:

$$\|\nabla f_i(D)\|_F$$

$$\leq \frac{1}{t} \sum_{i=1}^{t} \|Dv_i + e_i - z_i\|_2 \cdot \|v_i\|_2$$

$$+ \frac{1}{t^3} \sum_{i=1}^{t} \left[ \left( \frac{1}{\lambda I_p} + \frac{1}{t} N_i \right)^{-1} \right]_2^2 \cdot \|z_i\|_2 \cdot \|D\|_F$$

$$+ \frac{\lambda_3}{t^3} \left[ \left( \frac{1}{t} I_p + \frac{\lambda_3}{t} N_i \right)^{-1} \right]_F^2 \cdot \|D\|_F.$$

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.

F.2. Proof of P-Donsker

Proposition 11. 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 (2.8). Then we have

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

**Proof.** Let us consider $\{\ell(z, D)\}$ as a set of measurable functions indexed by $D \in \mathcal{D}$. As we showed in Proposition 7, $\mathcal{D}$ is a compact set. Also, we have proved that $\ell(z, D)$ is uniformly Lipschitz over $D$ (Proposition 9). Thus, $\{\ell(z, D)\}$ is P-Donsker (see the definition in Lemma 4). Furthermore, as $\ell(z, D)$ is non-negative and uniformly bounded, so is $\ell^2(z, D)$. So we have $E[\sqrt{t} \ell^2(z, D)]$ being uniformly bounded. Since we have verified all the hypotheses in Lemma 4, we obtain the result that

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

\[\square\]
F.3. Proof of convergence of $g_t(D)$

Theorem 12 (Convergence of the surrogate function $g_t(D_t)$). The surrogate function $g_t(D_t)$ defined in (3.5) converges almost surely, where $D_t$ is the solution produced by Algorithm 1.

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 $u_t$:

$$u_{t+1} - u_t = g_{t+1}(D_{t+1}) - g_t(D_t) = g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + g_{t+1}(D_t) - g_t(D_t)$$

$$= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{1}{t+1} \ell(z_{t+1}, D_t) - g_t(D_t)$$

$$+ \left[ \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|_2^2 - \frac{\lambda_3}{2(t+1)} \|D_t - M_t\|_F^2 \right].$$

And

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

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

$$- \frac{\lambda_3}{t+1} \text{Tr} ( (D_t - M_t)^T z_{t+1} u_{t+1}^T)$$

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

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

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

where the first equality is derived by the fact that $M_{t+1} = M_t + z_{t+1} u_{t+1}^T$, and the second equality is derived by the closed form solution of $u_{t+1}$ (see (3.4)).

Combining (F.14) and (F.15), we know that

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

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

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

$$- \|u_{t+1}\|_2^2 \right)$$

$$= \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|_2^2 \|u_{t+1}\|_2^2 - \frac{1}{2} \|u_{t+1}\|_2^2 \right) \leq 0.$$

Therefore,

$$u_{t+1} - u_t \leq g_{t+1}(D_{t+1}) - g_t(D_t) + \frac{1}{t+1} \ell(z_{t+1}, D_t)$$

$$- \frac{1}{t+1} g_t(D_t)$$

$$= g_{t+1}(D_{t+1}) - g_t(D_t) + \frac{f_t(D_t) - g_t(D_t)}{t+1}$$

$$+ \ell(z_{t+1}, D_t) - f_t(D_t)$$

$$\leq \ell(z_{t+1}, D_t) - f_t(D_t),$$

where $f_t(D)$ is defined in Proposition 11, and the last inequality holds because $D_{t+1}$ is the minimizer of $g_{t+1}(D)$ and $g_t(D)$ is a surrogate function of $f_t(D)$.

Let $\mathcal{F}_t$ be the filtration of the past information. We take the
From Proposition 11, 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
\[
\sum_{t=1}^{\infty} \mathbb{E}[\delta_t(u_{t+1} - u_t)] = \sum_{t \in \mathcal{T}} \mathbb{E}[u_{t+1} - u_t]
= \sum_{t \in \mathcal{T}} \mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]]
= \sum_{t=1}^{\infty} \mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]^{+}] \\ < +\infty, \ a.s. \quad (F.16)
\]

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

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

\[\quad \quad (F.17)\]
So the Frobenius norm of $\nabla G_t(D)$ follows immediately:

$$
\| \nabla G_t(D) \|_F \\
\leq \frac{1}{t} \left[ \| D \|_F \left( \lambda_1 \frac{A_{t+1}}{t+1} \right) + \lambda_1 \| v_{t+1} v_{t+1}^\top \|_F \\
\lambda_3 \frac{1}{t+1} \| I_d \|_F \right) + \lambda_1 \| (z_{t+1} - e_{t+1}) v_{t+1}^\top \|_F \\
+ \lambda_1 \frac{B_{t+1}}{t+1} \|_F + \lambda_3 \| z_{t+1} u_{t+1}^\top \|_F \\
+ \frac{\lambda_3}{t(t+1)} \| M_{t+1} \|_F \right] \nabla G_t(D) \|_F \\
= \frac{1}{t} \left[ \| D \|_F \left( \lambda_1 \frac{A_{t+1}}{t+1} \right) + \lambda_1 \| v_{t+1} v_{t+1}^\top \|_F \\
\lambda_1 \| (z_{t+1} - e_{t+1}) v_{t+1}^\top \|_F \\
+ \lambda_1 \frac{B_{t+1}}{t+1} \|_F + \lambda_3 \| z_{t+1} u_{t+1}^\top \|_F \\
+ \frac{\lambda_3}{t(t+1)} \| M_{t+1} \|_F \right].
$$

We know from Proposition 7 that all the terms in the above equation are uniformly bounded. Thus, there exist constants $c_1$, $c_2$ and $c_3$, such that

$$
\| \nabla G_t(D) \|_F \leq \frac{1}{t} \left[ c_1 \| D \|_F + c_2 \right] + \frac{c_3}{t(t+1)}.
$$

According to the first order Taylor expansion,

$$
G_t(D_{t+1}) - G_t(D_t) \\
= \text{Tr} \left( (D_{t+1} - D_t) \top \nabla G_t (\alpha D_t + (1 - \alpha) D_{t+1}) \right) \\
\leq \| D_{t+1} - D_t \|_F \cdot \| \nabla G_t (\alpha D_t + (1 - \alpha) D_{t+1}) \|_F,
$$

where $\alpha$ is a constant between 0 and 1. According to Proposition 7, $D_t$ 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 in

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

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

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

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

**Theorem 14** (Convergence of $f_t(D_t)$). Let $f_t(D_t)$ be the empirical loss function defined in (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 11 and Theorem 12 respectively. Then,

$$
b_t = g_t(D_t) - f_t(D_t)
= g_t(D_t) - f_t(D_t) + \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \| u_i \|_2^2 + \frac{\lambda_3}{2t} \| D_t - M_t \|_F^2
- \frac{1}{t^3} \sum_{i=1}^{t} \frac{1}{2} \left( \frac{1}{\lambda t^3} + \frac{\lambda_3}{t} \right) I_{t+1} \| D_t \|_F^2
- \frac{\lambda_3}{2t^3} \| D_t - M_t \|_F^2,
$$

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

$$
b_t = g_t(D_t) - f_t(D_t) = g_t(D_t) - f_t(D_t) + \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \| u_i \|_2^2 + \frac{\lambda_3}{2t} \| D_t - M_t \|_F^2
+ \frac{1}{t^3} \sum_{i=1}^{t} \frac{1}{2} \left( \frac{1}{\lambda t^3} + \frac{\lambda_3}{t} \right) I_{t+1} \| D_t \|_F^2
+ \frac{\lambda_3}{2t^3} \| D_t - M_t \|_F^2.
$$

Note that we naturally have

$$
\frac{q_t(D_t)}{t+1} \leq \frac{1}{t+1} \sum_{i=1}^{t} \frac{1}{2} \| u_i \|_2^2 + \frac{\lambda_3}{2t^3} \| D_t - M_t \|_F^2
\leq \frac{1}{t+1} \sum_{i=1}^{t} \frac{1}{2} \| u_i \|_2^2 + \frac{c}{2t(t+1)},
$$

where the second inequality holds as $D_t$ and $M_t$ are both uniformly bounded (see Proposition 7).
Also, from (F.14), we know
\[
\frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|_2^2 - \frac{1}{t} \sum_{i=1}^{t} \|u_i\|_2^2 = \frac{-1}{t(t+1)} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|_2^2 + \frac{1}{2(t+1)} \|u_{t+1}\|_2^2.
\]
And from (F.15)
\[
\frac{\lambda_3}{2(t+1)} \|D_t - M_t\|_F^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \\
\leq \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|_2^2 \|u_{t+1}\|_2^2 - \|u_{t+1}\|_2^2 \right).
\]
Thus,
\[
q_t(D_t) + \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|_2^2 \\
+ \frac{\lambda_3}{2(t+1)} \|D_t - M_t\|_F^2 \\
- \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|_2^2 - \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \\
\leq \frac{c}{2t(t+1)} + \frac{1}{2(t+1)} \|u_{t+1}\|_2^2 \\
+ \frac{1}{t+1} \left( -\frac{\lambda_3}{2} \|z_{t+1}\|_2^2 \|u_{t+1}\|_2^2 - \|u_{t+1}\|_2^2 \right) \\
= \frac{c}{2t(t+1)} - \frac{1}{2(t+1)} \|u_{t+1}\|_2^2 \\
- \frac{\lambda_3}{2(t+1)} \|z_{t+1}\|_2^2 \|u_{t+1}\|_2^2 \\
\leq \frac{c}{2(t+1)}.
\]
Therefore,
\[
\frac{b_t}{t+1} \\
\leq g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{\ell(z_{t+1}, D_t) - f'_t(D_t)}{t+1} \\
+ u_t - u_{t+1} + \frac{c}{2(t+1)} \\
\leq \frac{\ell(z_{t+1}, D_t) - f'_t(D_t)}{t+1} + u_t - u_{t+1} + \frac{c}{2(t+1)}.
\]
By taking the expectation conditioned on the past information \(F_t\), we have
\[
\frac{b_t}{t+1} \leq f(D_t) - f_t(D_t) + \mathbb{E}[u_t - u_{t+1} | F_t] + \frac{c}{2(t+1)} \\
\leq \frac{c_1}{\sqrt{t+1}} + |\mathbb{E}[u_t - u_{t+1} | F_t]| + \frac{c}{2(t+1)},
\]
where the second inequality holds by applying Proposition 11. Thus,
\[
\sum_{t=1}^{\infty} \frac{b_t}{t+1} \\
\leq \sum_{t=1}^{\infty} \frac{c_1}{\sqrt{t+1}} + \sum_{t=1}^{\infty} |\mathbb{E}[u_t - u_{t+1} | F_t]| \\
+ \sum_{t=1}^{\infty} \frac{c}{2(t+1)} \\
< + \infty.
\]
Here, the last inequality is derived by applying (F.16).

Next, we examine the difference between \(b_{t+1} - b_t\):
\[
|b_{t+1} - b_t| \\
= \bigg| g_{t+1}(D_{t+1}) - f_{t+1}(D_{t+1}) - g_{t}(D_t) + f_t(D_t) \bigg| \\
\leq \bigg| g_{t+1}(D_{t+1}) - g_{t+1}(D_{t+1} - g_t(D_t)) \bigg| \\
+ \bigg| g_{t+1}(D_{t+1}) - g_t(D_{t+1}) \bigg| \\
+ \bigg| g_t(D_{t+1}) - f_t(D_{t+1}) \bigg| \\
+ \bigg| f_t(D_{t+1}) - f_t(D_t) \bigg|.
\]
For the first term on the right hand side,
\[
\frac{c}{2t(t+1)} \\
\leq \bigg| g_{t+1}(D_{t+1}) - g_t(D_{t+1}) \bigg| \\
= \bigg| g_{t+1}(D_{t+1}) - g'_t(D_{t+1}) + \frac{1}{t+1} \sum_{i=1}^{t+1} \frac{1}{2} \|u_i\|_2^2 \\
- \frac{1}{t} \sum_{i=1}^{t} \frac{1}{2} \|u_i\|_2^2 + \frac{\lambda_3}{2t} \|D_t - M_t\|_F^2 \\
- \frac{\lambda_3}{2(t+1)} \|D_{t+1} - M_{t+1}\|_F^2 \\
- \frac{\lambda_3}{2(t+1)} \|z_{t+1}\|_2^2 \|u_{t+1}\|_2^2 \\
\leq \frac{c_1}{\sqrt{t+1}} + \mathbb{E}[u_t - u_{t+1} | F_t] + \frac{c}{2(t+1)}.
\]
By taking the expectation conditioned on the past information \(F_t\), we have
\[
\frac{b_t}{t+1} \leq \frac{1}{t+1} \ell(z_{t+1}, D_{t+1}) - \frac{1}{t+1} g'_t(D_{t+1}) + \frac{c_1}{t+1} \\
\leq \frac{c_2}{t+1},
\]
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 7), and \( \zeta_2 \) holds because \( \ell(z_{t+1}, D_{t+1}) \) and \( g'_t(D_{t+1}) \) are uniformly bounded (see Corollary 8).

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

\[
\left| f_t(D_{t+1}) - f_t(D_t) \right| \\
\leq \frac{1}{t+1} \ell(z_{t+1}, D_{t+1}) - \frac{1}{t+1} f'_t(D_{t+1}) + \frac{c_3}{t+1} \\
\leq \frac{c_4}{t+1},
\]

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 10).

From Corollary 8 and Corollary 10, 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

\[
\left| g_t(D_{t+1}) - g_t(D_t) \right| \leq \kappa_1 \left\| D_{t+1} - D_t \right\|_F \leq \frac{\kappa_3}{t+1}, \\
\left| f_t(D_{t+1}) - f_t(D_t) \right| \leq \kappa_2 \left\| D_{t+1} - D_t \right\|_F \leq \frac{\kappa_4}{t+1}.
\]

Here, \( \zeta_4 \) and \( \zeta_5 \) are derived by applying Proposition 13 and \( \kappa_3 \) and \( \kappa_4 \) are some uniform constants.

Finally, we have a bound for (F.20):

\[
|b_{t+1} - b_t| \leq \frac{\kappa_0}{t+1},
\]

where \( \kappa_0 \) is some uniform constant.

By applying Lemma 6, 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 12 that \( g_t(D_t) \) converges almost surely, we conclude that \( f_t(D_t) \) converges almost surely to the same limit of \( g_t(D_t) \) (or, \( g_t(D_t) \)).

**Theorem 15 (Convergence of \( f(D_t) \)).** Let \( f(D_t) \) be the expected loss function we defined in (2.8) and let \( D_t \) be the optimal solution produced by Algorithm 1. Then \( f(D_t) \) converges almost surely to the same limit of \( f_t(D_t) \) (or, \( g_t(D_t) \)).

**Proof.** According to the central limit theorem, we know that \( \sqrt{t}(f(D_t) - f(D_t)) \) is bounded, implying

\[
\lim_{t \to +\infty} f(D_t) - f_t(D_t) = 0, \quad \text{a.s.}
\]

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

**Proposition 16 (Gradient of \( f(D) \)).** Let \( f(D) \) be the expected loss function which is defined in (2.8). Then, \( f(D) \) is continuously differentiable and \( \nabla f(D) = \mathbb{E}_z[\nabla \ell(z, D)] \). Moreover, \( \nabla f(D) \) is uniformly Lipschitz on \( D \).

**Proof.** We have shown in Proposition 9 that \( \ell(z, D) \) is continuously differentiable, \( f(D) \) is also continuously differentiable and we have \( \nabla f(D) = \mathbb{E}_z[\nabla \ell(z, D)] \).

Next, we prove the Lipschitz of \( \nabla f(D) \). Let \( v'(z', D') \) and \( e'(z', D') \) be the minimizer of \( \ell(z', D', v, e) \). Since \( \ell(z, D, v, e) \) 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

\[
\frac{\partial \ell(z, D, v, e)}{\partial e} = 0,
\]

which implies

\[
\lambda_1(z - Dv - e) \in \lambda_2 \|e\|_1.
\]

Hence,

\[
|(z - Dv - e)_j| = \frac{\lambda_2}{\lambda_1}, \quad \forall j \in A.
\]

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 \( |(z'' - D''v'' - e'')_j| < \frac{\lambda_2}{\lambda_1} \) and \( e''_j = 0 \). That is, the support set of \( e' \) will not change.

Let us denote \( H = [D I_p], r = [v^T e^T]^T \) and define the function

\[
\tilde{\ell}(z, H, r_A) = \frac{\lambda_1}{2} \|z - H_A r_A\|_2^2 + \frac{1}{2} \|I_0 r_A\|_2^2 + \lambda_2 \|I_0 r_A\|_1.
\]

Since \( \tilde{\ell}(z, D_A, r_A) \) is strongly convex, there exists a uniform constant \( \kappa_1 \), such that for all \( \|r''_A\|_2 \),

\[
\tilde{\ell}(z', H'_A, r''_A) - \tilde{\ell}(z', H'_A, r'_A) \\
\geq \kappa_1 \|r''_A - r'_A\|_2^2 \\
= \kappa_1 \left( \|v'' - v'||_2^2 + \|e'' - e'_A\|_2^2 \right).
\]

On the other hand,

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

where \( \delta_{t+1} \) converges almost surely to the same limit of \( f_t(D_t) \) (or, \( g_t(D_t) \)).
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 with Lipschitz constant \( 1 \) and \( z'' \) are all uniformly bounded. Hence, there exists uniform constants \( c_1 \) and \( c_2 \), such that

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

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

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

Therefore, \( r(z, D) \) is Lipschitz and so are \( v(z, D) \) and \( e(z, D) \). Note that according to Proposition 9,

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

Thus,

\[
\left\| \nabla f(D') - \nabla f(D'') \right\|_F \\
\leq \mathbb{E}_z \left[ \left\| H' r' \right\|_2 \left\| v' - v'' \right\|_2 + \left\| H' - H'' \right\|_F \left\| r' v''^\top \right\|_F \right] \\
+ \left\| H'' \left\|_2 \right\| r' - r'' \right\|_2 \left\| v'' \right\|_2 + \left\| z' \right\|_2 \left\| v' - v'' \right\|_2 \] \\
\leq \mathbb{E}_z \left[ (\gamma_1 + \gamma_2 \left\| z' \right\|_2) \left\| H' - H'' \right\|_F \right] \\
\leq \gamma_0 \left\| D' - D'' \right\|_F,
\]

where \( \gamma_0 \), \( \gamma_1 \) and \( \gamma_2 \) are all uniform constants. Here, \( \zeta_1 \) holds because for any function \( s(z) \), we have \( \mathbb{E}_z \left[ \left\| s(z) \right\|_F \right] \leq \mathbb{E}_z \left[ \left\| s(z) \right\|_F \right] \). \( \zeta_2 \) 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 \( z \) is uniformly bounded and actually \( \left\| H' - H'' \right\|_F = \left\| D' - D'' \right\|_F \). Thus, we complete the proof.

---

**F.7. Proof of stationary point**

**Theorem 17** (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 (2.8). Then \( D_t \) converges to a stationary point of \( f(D) \) when \( t \) goes to infinity.

**Proof.** Since \( \frac{1}{T} A_t \) and \( \frac{1}{T} B_t \) are uniformly bounded (Proposition 7), there exist sub-sequences of \( \left\{ \frac{1}{T} A_t \right\} \) and \( \left\{ \frac{1}{T} B_t \right\} \) 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

\[
g_t(D_t + h_k W) \geq f_t(D_t + h_k W).
\]

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

\[
g_\infty(D_\infty + h_k W) \geq f_\infty(D_\infty + h_k W).
\]

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,

\[
\text{Tr}(h_k W^\top \nabla g_\infty(D_\infty)) + o(h_k W)
\geq \text{Tr}(h_k W^\top \nabla f(D_\infty)) + o(h_k W)
\]

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

\[
\text{Tr} \left( \frac{1}{\left\| W \right\|_F} W^\top \nabla g_\infty(D_\infty) \right) + \frac{o(h_k W)}{h_k \left\| W \right\|_F}
\geq \text{Tr} \left( \frac{1}{\left\| W \right\|_F} W^\top \nabla f(D_\infty) \right) + \frac{o(h_k W)}{h_k \left\| W \right\|_F}
\]

Now let \( k \) go to infinity.

\[
\text{Tr} \left( \frac{1}{\left\| W \right\|_F} W^\top \nabla g_\infty(D_\infty) \right) \geq \text{Tr} \left( \frac{1}{\left\| W \right\|_F} W^\top \nabla f(D_\infty) \right).
\]

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


Eriksson, Brian, Balzano, Laura, and Nowak, Robert D. High-rank matrix completion and subspace clustering with missing data. CoRR, abs/1112.5629, 2011.


Online Low-Rank Subspace Clustering by Basis Dictionary Pursuit


