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:

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

which implies

$$\begin{aligned} U^* &= \left(\frac{1}{\lambda_3} I_p + Y^\top Y\right)^{-1} Y^\top D \\ &= \lambda_3 \sum_{j=0}^{+\infty} (-\lambda_3 Y^\top Y)^j Y^\top D \\ &= \lambda_3 Y^\top \Big[\sum_{j=0}^{+\infty} (-\lambda_3 Y Y^\top)^j\Big] D \\ &= Y^\top \left(\frac{1}{\lambda_3} I_p + Y Y^\top\right)^{-1} D. \end{aligned}$$

Then, its transpose is given by

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

Note that $\boldsymbol{u}_i$ is the column of $U^\top$. So for each $i \in [n]$,

$$\begin{aligned} \boldsymbol{u}_i^* &= D^\top \left(\frac{1}{\lambda_3} I_p + \sum_{i=1}^n \boldsymbol{y}_i \boldsymbol{y}_i^\top\right)^{-1} \boldsymbol{y}_i \\ &= \frac{1}{n} D^\top \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n} N_n\right)^{-1} \boldsymbol{y}_i, \end{aligned}$$

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

Also, we have

$$\begin{aligned} Y U^{*\top} &= Y \left(\frac{1}{\lambda_3} I_p + Y^\top Y\right)^{-1} Y^\top D \\ &= \lambda_3 Y \left(I_p + \lambda_3 Y^\top Y\right)^{-1} Y^\top D \\ &= \lambda_3 Y \Big[\sum_{j=0}^{+\infty} (-\lambda_3 Y^\top Y)^j\Big] Y^\top D \\ &= \lambda_3 \sum_{j=0}^{+\infty} (-\lambda_3)^j (Y Y^\top)^{j+1} D \\ &= D - \left(I_p + \lambda_3 Y Y^\top\right)^{-1} D \\ &= D - \left(I_p + \lambda_3 \sum_{i=1}^n \boldsymbol{y}_i \boldsymbol{y}_i^\top\right)^{-1} D \\ &= D - \frac{1}{n}\left(\frac{1}{n} I_p + \frac{\lambda_3}{n} N_n\right)^{-1} D. \end{aligned}$$

Thus,

$$\begin{aligned} h(Y,D) = & \sum_{i=1}^n \frac{1}{2} \left\| \frac{1}{n} D^\top \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n} N_n\right)^{-1} \boldsymbol{y}_i \right\|_2^2 \\ & + \frac{\lambda_3}{2} \left\| \frac{1}{n}\left(\frac{1}{n} I_p + \frac{\lambda_3}{n} N_n\right)^{-1} D \right\|_F^2 \\ = & \frac{1}{n^2} \sum_{i=1}^n \frac{1}{2} \left\| D^\top \left(\frac{1}{\lambda_3 n} I_p + \frac{1}{n} N_n\right)^{-1} \boldsymbol{y}_i \right\|_2^2 \\ & + \frac{\lambda_3}{2n^2} \left\| \left(\frac{1}{n} I_p + \frac{\lambda_3}{n} N_n\right)^{-1} D \right\|_F^2. \end{aligned}$$

C. Algorithm Details

Algorithm 2 Solving $\boldsymbol{v}$ and $\boldsymbol{e}$

Require: $D \in \mathbb{R}^{p \times d}$, $\boldsymbol{z} \in \mathbb{R}^p$, parameters λ_1 and λ_2
Ensure: Optimal $\boldsymbol{v}$ and $\boldsymbol{e}$.
1: Set $\boldsymbol{e} = \boldsymbol{0}$.
2: **repeat**
3: Update $\boldsymbol{v}$:

$$\boldsymbol{v} = (D^\top D + \frac{1}{\lambda_1} I)^{-1} D^\top (\boldsymbol{z} - \boldsymbol{e}).$$

4: Update $\boldsymbol{e}$:

$$\boldsymbol{e} = \mathcal{S}_{\lambda_2/\lambda_1}[\boldsymbol{z} - D\boldsymbol{v}].$$

5: **until** convergence

For Algorithm 2, we set a threshold $\epsilon = 10^{-3}$. Let $\{\boldsymbol{v}', \boldsymbol{e}'\}$ and $\{\boldsymbol{v}'', \boldsymbol{e}''\}$ be the two consecutive iterates. If the maximum of $\|\boldsymbol{v}' - \boldsymbol{v}''\|_2 / \|\boldsymbol{v}'\|_2$ and $\|\boldsymbol{e}' - \boldsymbol{e}''\|_2 / \|\boldsymbol{e}'\|_2$ is less than ϵ, then we stop Algorithm 2.

Algorithm 3 Solving D

Require: $D \in \mathbb{R}^{p \times d}$ in the previous iteration, accumulation matrix M, A and B, parameters λ_1 and λ_3.
Ensure: Optimal D (updated).
1: Denote $\widehat{A} = \lambda_1 A + \lambda_3 I$ and $\widehat{B} = \lambda_1 B + \lambda_3 M$.
2: **repeat**
3: **for** $j = 1$ to d **do**
4: Update the jth column of D:

$$\boldsymbol{d}_j \leftarrow \boldsymbol{d}_j - \frac{1}{\widehat{A}_{jj}} \left(D\widehat{\boldsymbol{a}}_j - \widehat{\boldsymbol{b}}_j \right)$$

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).**

	OLRSC	ORPCA	LRR	LRR2	SSC
Mushrooms	**85.09** 8.78	65.26 8.30	58.44 46.82	56.38 8.55	54.16 32 min
DNA	**67.11** 2.58	53.11 2.09	44.01 23.67	45.32 1.65	52.23 3 min
Protein	43.30 24.66	40.22 22.90	40.31 921.58	40.00 98.33	**44.27** 65 min
USPS	**65.95** 33.93	55.70 27.01	52.98 257.25	58.69 71.15	47.58 50 min
MNIST-7K	**58.04** 42.99	55.40 39.84	54.77 512.37	54.27 95.21	45.56 26 min
MNIST-10K	**56.79** 67	54.66 56	55.15 24 min	53.67 153	44.90 84 min
MNIST-20K	**57.74** 129	54.10 121	55.23 32 min	54.53 360	43.91 7 hours

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

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 Θ of $\mathbb{R}^d$. Suppose that there exists a constant K such that*

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

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

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

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

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

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

Lemma 5 (Sufficient condition of convergence for a stochastic process (Bottou, 1998))**.** *Let $(\Omega, \mathcal{F}, P)$ 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*

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

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,

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

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 $\{\boldsymbol{u}_t\}$, $\{\boldsymbol{v}_t\}$, $\{\boldsymbol{e}_t\}$ and $\{D_t\}$ be the optimal solutions produced by Algorithm 1. Then,*

1. *$\boldsymbol{v}_t$, $\boldsymbol{e}_t$, $\frac{1}{t}A_t$ 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. *$\boldsymbol{u}_t$ is uniformly bounded.*

Proof. Let us consider the optimization problem of solving $\boldsymbol{v}$ and $\boldsymbol{e}$. As the trivial solution $\{\boldsymbol{v}_t', \boldsymbol{e}_t'\} = \{\boldsymbol{0}, \boldsymbol{z}_t\}$ are feasible, we have

$$\tilde{\ell}_1(\boldsymbol{z}_t, D_{t-1}, \boldsymbol{v}_t', \boldsymbol{e}_t') = \lambda_2 \left\| \boldsymbol{z}_t \right\|_1 .$$

Therefore, the optimal solution should satisfy:

$$\frac{\lambda_1}{2} \left\| \boldsymbol{z}_t - D_{t-1}\boldsymbol{v}_t - \boldsymbol{e}_t \right\|_2^2 + \frac{1}{2} \left\| \boldsymbol{v}_t \right\|_2^2 + \lambda_2 \left\| \boldsymbol{e}_t \right\|_1 \leq \lambda_2 \left\| \boldsymbol{z}_t \right\|_1 ,$$

which implies

$$\frac{1}{2} \left\| \boldsymbol{v}_t \right\|_2^2 \leq \lambda_2 \left\| \boldsymbol{z}_t \right\|_1 ,$$
$$\lambda_2 \left\| \boldsymbol{e}_t \right\|_1 \leq \lambda_2 \left\| \boldsymbol{z}_t \right\|_1 .$$

Since $\boldsymbol{z}_t$ is uniformly bounded (Assumption 1), $\boldsymbol{v}_t$ and $\boldsymbol{e}_t$ are uniformly bounded.

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

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

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

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

1. $M_t = \sum_{i=1}^{t} \boldsymbol{y}_i \boldsymbol{u}_i^\top$ (definition of M_t).

2. $\boldsymbol{u}_t = (\|\boldsymbol{y}_t\|_2^2 + \frac{1}{\lambda_3})^{-1}(D_{t-1} - M_{t-1})^\top \boldsymbol{y}_t$ (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:

$$D_t = (\lambda_1 B_t + \lambda_3 M_t)Q_t^{-1}, \tag{F.1}$$

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

$$\begin{aligned}
& M_{t+1} - M_t \\
&= \boldsymbol{y}_{t+1}\boldsymbol{u}_{t+1}^\top \\
&= \left(\left\|\boldsymbol{y}_{t+1}\right\|_2^2 + \frac{1}{\lambda_3} \right)^{-1} \boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^\top (D_t - M_t) \\
&= P_t D_t - P_t M_t \\
&= P_t(\lambda_1 B_t + \lambda_3 M_t)Q_t^{-1} - P_t M_t,
\end{aligned} \tag{F.2}$$

where

$$P_t = \left(\left\|\boldsymbol{y}_{t+1}\right\|_2^2 + \frac{1}{\lambda_3} \right)^{-1} \boldsymbol{y}_{t+1}\boldsymbol{y}_{t+1}^\top.$$

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

$$M_{t+1} = \left(M_t - \lambda_1 P_t M_t A_t Q_t^{-1}\right) + \lambda_1 P_t B_t Q_t^{-1}. \tag{F.3}$$

By applying the Taylor expansion on Q_t^{-1}, we have

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

Thus,

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

So M_{t+1} is given by

$$\begin{aligned}
M_{t+1} = {} & (I_d - P_t)M_t \\
& \underbrace{+ P_t M_t \left(I_d + \frac{\lambda_1}{\lambda_3}A_t\right)^{-1} + \lambda_1 P_t B_t Q_t^{-1}}_{W_t}.
\end{aligned} \tag{F.4}$$

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

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

Since we assume that $\{\boldsymbol{z}_t\}$ are uniformly upper bounded (Assumption 1), there exists a constant α_1, such that for all $t > 0$,

$$\|\boldsymbol{z}_t\|_2 \leq \alpha_1.$$

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,

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

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:

$$\begin{aligned} \|W_t\| &\leq \|P_t\| \cdot \|M_t\| \cdot \left\| \left(I_d + \frac{\lambda_1}{\lambda_3} A_t\right)^{-1} \right\| + c_2 \\ &\overset{\zeta_1}{\leq} \frac{\alpha_1^2}{\alpha_1^2 + \frac{1}{\lambda_3}} \cdot \frac{\lambda_3}{\lambda_3 + \lambda_1 \beta_0 t} \|M_t\| + c_2 \\ &\leq \frac{c_3}{t} \|M_t\| + c_2, \end{aligned} \tag{F.6}$$

where ζ_1 is derived by utilizing the assumption that $\boldsymbol{z}$ is upper bounded by α_1 and the smallest singular value of $\frac{1}{t} A_t$ is lower bounded by β_0. The last inequality always holds for some uniform constant c_3.

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

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

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

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

$$\begin{aligned} \|W_t\| &\leq w^*, \\ \frac{c_3}{t} m_1 + 0.5 w^* + c_2 &\leq w^*. \end{aligned} \tag{F.8}$$

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,

$$\begin{aligned} &\|M_{\tau+1}\| \\ = &\left\| \prod_{i=1}^{\tau} (I_p - P_i) M_1 + \sum_{j=1}^{\tau-1} \prod_{i=j+1}^{\tau} (I_p - P_i) W_j + W_\tau \right\| \\ \leq &\left\| \prod_{i=1}^{\tau} (I_p - P_i) M_1 \right\| + \left\| \sum_{j=1}^{\tau-1} \prod_{i=j+1}^{\tau} (I_p - P_i) W_j + W_\tau \right\| \\ \overset{\zeta_1}{\leq} &\left\| \prod_{i=1}^{\tau} (I_p - P_i) \right\| \cdot \|M_1\| + \tau w^* \\ \overset{\zeta_2}{\leq} &(1 - \alpha_0) m_1 + \tau w^*. \end{aligned}$$

Here, ζ_1 holds because $\left\| \prod_{i=j+1}^{\tau} (I_p - P_i) \right\| \leq 1$ for all $j \in [\tau - 1]$. ζ_2 holds because the singular values of P_i's have span over the diagonal so the largest singular value of $\prod_{i=1}^{\tau} (I_p - P_i)$ is $1 - \alpha_0$, where α_0 is the lower bound for all $\boldsymbol{z}_i$'s (see Assumption 1).

For $M_{2(\tau+1)}$, we can similarly obtain

$$\left\| M_{2(\tau+1)} \right\| \leq (1 - \alpha_0)^2 m_1 + (1 - \alpha_0) \tau w^* + \tau w^*.$$

More generally, for any integer $k \le k^*$,

$$\begin{aligned}\left\|M_{k(\tau+1)}\right\| \le&(1-\alpha_0)^k m_1 + \sum_{j=0}^{k-1}(1-\alpha_0)^j \tau w^* \\ \le& m_1 + \frac{\tau w^*}{\alpha_0}.\end{aligned}$$

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

$$\begin{aligned}\|M_i\| &\le m_1 + \frac{\tau w^*}{\alpha_0} + (i-(k-1)(\tau+1))w^* \\ &\le m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*.\end{aligned}$$

Therefore, for all the current M_t's, we can bound them as follows:

$$\|M_t\| \le m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*,\ \forall\, t = 1, 2, \cdots, t^*. \tag{F.9}$$

From (F.8) and (F.9), we know that for all $t \le t^*$,

$$\begin{aligned}&\|W_t\| \le w^*, \\ &\|M_t\| \le m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*.\end{aligned}$$

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,

$$\begin{aligned}\|M_{t^*+1}\| &\le \left\|M_{k^*(\tau+1)}\right\| + (t^*+1-k^*(\tau+1))w^* \\ &\le m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*. \end{aligned}\tag{F.10}$$

For $\|W_{t^*+1}\|$, from (F.6), we know

$$\begin{aligned}\|W_{t^*+1}\| &\le \frac{c_3}{t^*+1}\|M_{t^*+1}\| + c_2 \\ &\le \frac{c_3}{t^*+1}(m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*) + c_2 \\ &= \frac{c_3 m_1}{t^*+1} + \frac{c_3\tau}{t^*+1}(\frac{1}{\alpha_0}+1)w^* + c_2 \\ &\overset{\zeta_1}{\le} \frac{c_3 m_1}{t^*+1} + 0.5w^* + c_2 \\ &\overset{\zeta_2}{\le} w^*. \end{aligned}\tag{F.11}$$

Here, ζ_1 is derived by utilizing (F.7) and ζ_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 $t \le t^*$, still holds for $t = t^* + 1$. Thus we complete the induction and conclude that for all $t > 0$, we have

$$\begin{aligned}&\|M_t\| \le m_1 + \frac{\tau w^*}{\alpha_0} + \tau w^*, \\ &\|W_t\| \le w^*.\end{aligned}$$

Thus, M_t is uniformly bounded.

From (F.1), we know that

$$\begin{aligned}D_t &= \lambda_1 B_t\left(\lambda_1 A_t + \lambda_3 I_d\right)^{-1} + \lambda_3 M_t\left(\lambda_1 A_t + \lambda_3 I_d\right)^{-1} \\ &= \lambda_1\left(\frac{1}{t}B_t\right)\left(\frac{\lambda_1}{t}A_t + \frac{\lambda_3}{t}I_d\right)^{-1} \\ &\quad + \frac{\lambda_3}{t}M_t\left(\frac{\lambda_1}{t}A_t + \frac{\lambda_3}{t}I_d\right)^{-1}.\end{aligned}$$

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 $\boldsymbol{u}_t$, and note that we have proved the uniform boundedness of D_t and M_t, we conclude that $\{\boldsymbol{u}_t\}$ are uniformly bounded. □

Corollary 8. *Let $\boldsymbol{v}_t$, $\boldsymbol{e}_t$, $\boldsymbol{u}_t$ and D_t be the optimal solutions produced by Algorithm 1.*

1. *$\tilde{\ell}(\boldsymbol{z}_t, D_t, \boldsymbol{v}_t, \boldsymbol{e}_t)$ and $\ell(\boldsymbol{z}_t, D_t)$ 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 $\tilde{\ell}(\boldsymbol{z}_t, D_t, \boldsymbol{v}_t, \boldsymbol{e}_t)$ and notice that $\boldsymbol{z}_t$, D_t, $\boldsymbol{v}_t$ and $\boldsymbol{e}_t$ are all uniformly bounded. This implies that $\tilde{\ell}(\boldsymbol{z}_t, D_t, \boldsymbol{v}_t, \boldsymbol{e}_t)$ is uniformly bounded and so is $\ell(\boldsymbol{z}_t, D_t)$. 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 $\tilde{\ell}(\boldsymbol{z}_t, D_t, \boldsymbol{v}_t, \boldsymbol{e}_t)$ 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 $D \in \mathcal{D}$.

$$\begin{aligned}\|\nabla g_t(D)\|_F &= \left\|\lambda_1 D\left(\frac{A_t}{t} + \frac{\lambda_3}{t}I_d\right) - \lambda_1\frac{B_t}{t} - \frac{\lambda_3}{t}M_t\right\|_F \\ &\le \lambda_1\|D\|_F\left(\left\|\frac{A_t}{t}\right\|_F + \left\|\frac{\lambda_3}{t}I_d\right\|_F\right) \\ &\quad + \lambda_1\left\|\frac{B_t}{t}\right\|_F + \left\|\frac{\lambda_3}{t}M_t\right\|_F.\end{aligned}$$

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 $\tilde{\ell}(\boldsymbol{z}, D, \boldsymbol{v}, \boldsymbol{e})$ as:*

$$\{\boldsymbol{v}', \boldsymbol{e}'\} = \operatorname*{arg\,min}_{\boldsymbol{v},\boldsymbol{e}} \tilde{\ell}(\boldsymbol{z}, D, \boldsymbol{v}, \boldsymbol{e}).$$

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

$$\nabla_D \ell(\boldsymbol{z}, D) = (D\boldsymbol{v}' + \boldsymbol{e}' - \boldsymbol{z})\boldsymbol{v}'^\top.$$

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

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

$$\begin{aligned} &\mathbb{R}^{d+p} \times \mathcal{D} \to \mathbb{R} \\ &([\boldsymbol{v};\ \boldsymbol{e}], D) \mapsto \tilde{\ell}(\boldsymbol{z}, D, \boldsymbol{v}, \boldsymbol{e}). \end{aligned}$$

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

$$\nabla_D \ell(\boldsymbol{z}, D) = (D\boldsymbol{v}' + \boldsymbol{e}' - \boldsymbol{z})\boldsymbol{v}'^\top.$$

Since every term in $\nabla_D \ell(\boldsymbol{z}, D)$ is uniformly bounded (Assumption 1 and Proposition 7), we conclude that the gradient of $\ell(\boldsymbol{z}, D)$ is uniformly bounded, implying that $\ell(\boldsymbol{z}, D)$ is uniformly Lipschitz w.r.t. D. ☐

Corollary 10. *Let $f_t(D)$ be the empirical loss function defined in (2.6). Then $f_t(D)$ is uniformly bounded and Lipschitz.*

Proof. Since $\ell(\boldsymbol{z}, L)$ 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:

$$\begin{aligned} \frac{1}{t}h(Z, D) = \frac{1}{t^3}\sum_{i=1}^{t} \frac{1}{2}\left\| D^\top \left(\frac{1}{\lambda_3 t} I_p + \frac{1}{t} N_t\right)^{-1} \boldsymbol{z}_i \right\|_2^2 \\ + \frac{\lambda_3}{2t^3}\left\| \left(\frac{1}{t} I_p + \frac{\lambda_3}{t} N_t\right)^{-1} D \right\|_F^2 \end{aligned}$$

where $N_t = \sum_{i=1}^{t} \boldsymbol{z}_i \boldsymbol{z}_i^\top$. Since every term in the above equation can be uniformly bounded, $h(Z, D)$ is uniformly bounded and so is $f_t(D)$.

To show that $f_t(D)$ is uniformly Lipschitz, we show that its gradient can be uniformly bounded.

$$\begin{aligned} &\nabla f_t(D) \\ &= \frac{1}{t}\sum_{i=1}^{t} \nabla \ell(\boldsymbol{z}_i, D) + \frac{1}{t}\nabla h(Z, D) \\ &= \frac{1}{t}\sum_{i=1}^{t} (D\boldsymbol{v}_i + \boldsymbol{e}_i - \boldsymbol{z}_i)\boldsymbol{v}_i^\top \\ &\quad + \frac{1}{t^3}\sum_{i=1}^{t} \left(\frac{1}{\lambda_3 t} I_p + \frac{1}{t} N_t\right)^{-1} \boldsymbol{z}_i \boldsymbol{z}_i^\top \left(\frac{1}{\lambda_3 t} I_p + \frac{1}{t} N_t\right)^{-1} D \\ &\quad + \frac{\lambda_3}{t^3}\left(\frac{1}{t} I_p + \frac{\lambda_3}{t} N_t\right)^{-2} D. \end{aligned}$$

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

$$\begin{aligned} &\left\|\nabla f_t(D)\right\|_F \\ &\le \frac{1}{t}\sum_{i=1}^{t} \left\|D\boldsymbol{v}_i + \boldsymbol{e}_i - \boldsymbol{z}_i\right\|_2 \cdot \left\|\boldsymbol{v}_i\right\|_2 \\ &\quad + \frac{1}{t^3}\sum_{i=1}^{t} \left\| \left(\frac{1}{\lambda_3 t} I_p + \frac{1}{t} N_t\right)^{-1} \right\|_F^2 \cdot \left\|\boldsymbol{z}_i\right\|_2^2 \cdot \left\|D\right\|_F \\ &\quad + \frac{\lambda_3}{t^3}\left\| \left(\frac{1}{t} I_p + \frac{\lambda_3}{t} N_t\right)^{-1} \right\|_F^2 \cdot \left\|D\right\|_F. \end{aligned}$$

One can easily check that the right side of the inequality is uniformly bounded. Thus $\left\|\nabla f_t(D)\right\|_F$ is uniformly bounded, implying that $f_t(D)$ is uniformly Lipschitz. ☐

F.2. Proof of P-Donsker

Proposition 11. *Let $f_t'(D) = \frac{1}{t}\sum_{i=1}^{t} \ell(\boldsymbol{z}_i, D)$ and $f(D)$ be the expected loss function defined in (2.8). Then we have*

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

Proof. Let us consider $\{\ell(\boldsymbol{z}, D)\}$ 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 $\ell(\boldsymbol{z}, D)$ is uniformly Lipschitz over D (Proposition 9). Thus, $\{\ell(\boldsymbol{z}, D)\}$ is P-Donsker (see the definition in Lemma 4). Furthermore, as $\ell(\boldsymbol{z}, D)$ is non-negative and uniformly bounded, so is $\ell^2(\boldsymbol{z}, D)$. So we have $\mathbb{E}_{\boldsymbol{z}}[\ell^2(\boldsymbol{z}, D)]$ being uniformly bounded. Since we have verified all the hypotheses in Lemma 4, we obtain the result that

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

☐

F.3. Proof of convergence of $g_t(D)$

Theorem 12 (Convergence of the surrogate function $g_t(D_t)$)**.** *The surrogate function $g_t(D_t)$ we 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:

$$\begin{aligned}
&u_{t+1} - u_t \\
&= g_{t+1}(D_{t+1}) - g_t(D_t) \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + g_{t+1}(D_t) - g_t(D_t) \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) \\
&\quad + \frac{1}{t+1}\ell(\boldsymbol{z}_{t+1}, D_t) - \frac{1}{t+1} g_t'(D_t) \\
&\quad + \Bigg[\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{\lambda_3}{2(t+1)}\|D_t - M_{t+1}\|_F^2 \\
&\quad - \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 - \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2\Bigg] \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{f_t'(D_t) - g_t'(D_t)}{t+1} \\
&\quad + \frac{\ell(\boldsymbol{z}_{t+1}, D_t) - f_t'(D_t)}{t+1} \\
&\quad + \Bigg[\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{\lambda_3}{2(t+1)}\|D_t - M_{t+1}\|_F^2 \\
&\quad - \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 - \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2\Bigg].
\end{aligned} \tag{F.12}$$

Here,

$$g_t'(D_t) = \frac{1}{t}\sum_{i=1}^{t}\tilde{\ell}(\boldsymbol{z}_i, D, \boldsymbol{v}_i, \boldsymbol{e}_i). \tag{F.13}$$

First, we bound the last four terms. We have

$$\begin{aligned}
&\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 - \frac{1}{t}\sum_{i=1}^{t}\|\boldsymbol{u}_i\|_2^2 \\
&= \frac{-1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{1}{2(t+1)}\|\boldsymbol{u}_{t+1}\|_2^2 \\
&\le \frac{1}{2(t+1)}\|\boldsymbol{u}_{t+1}\|_2^2.
\end{aligned} \tag{F.14}$$

And

$$\begin{aligned}
&\frac{\lambda_3}{2(t+1)}\|D_t - M_{t+1}\|_F^2 - \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2 \\
&= \frac{-\lambda_3}{2t(t+1)}\|D_t - M_t\|_F^2 + \frac{\lambda_3}{2(t+1)}\left\|\boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right\|_F^2 \\
&\quad - \frac{\lambda_3}{t+1}\operatorname{Tr}\left((D_t - M_t)^\top \boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right) \\
&= \frac{-\lambda_3}{2t(t+1)}\|D_t - M_t\|_F^2 + \frac{\lambda_3}{2(t+1)}\left\|\boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right\|_F^2 \\
&\quad - \frac{\lambda_3}{t+1}\left(\|\boldsymbol{z}_{t+1}\|_2^2 + \frac{1}{\lambda_3}\right)\|\boldsymbol{u}_{t+1}\|_2^2 \\
&\le \frac{1}{t+1}\left(\frac{\lambda_3}{2}\left\|\boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right\|_F^2 - (\lambda_3\|\boldsymbol{z}_{t+1}\|_2^2 + 1)\|\boldsymbol{u}_{t+1}\|_2^2\right) \\
&\le \frac{1}{t+1}\left(-\frac{\lambda_3}{2}\|\boldsymbol{z}_{t+1}\|_2^2\|\boldsymbol{u}_{t+1}\|_2^2 - \|\boldsymbol{u}_{t+1}\|_2^2\right),
\end{aligned} \tag{F.15}$$

where the first equality is derived by the fact that $M_{t+1} = M_t + \boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top$, and the second equality is derived by the closed form solution of $\boldsymbol{u}_{t+1}$ (see (3.4)).

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

$$\begin{aligned}
&\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 - \frac{1}{t}\sum_{i=1}^{t}\|\boldsymbol{u}_i\|_2^2 \\
&\quad + \frac{\lambda_3}{2(t+1)}\|D_t - M_{t+1}\|_F^2 - \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2 \\
&\le \frac{1}{2(t+1)}\|\boldsymbol{u}_{t+1}\|_2^2 + \frac{1}{t+1}\Big(-\frac{\lambda_3}{2}\|\boldsymbol{z}_{t+1}\|_2^2\|\boldsymbol{u}_{t+1}\|_2^2 \\
&\quad - \|\boldsymbol{u}_{t+1}\|_2^2\Big) \\
&= \frac{1}{t+1}\left(-\frac{\lambda_3}{2}\|\boldsymbol{z}_{t+1}\|_2^2\|\boldsymbol{u}_{t+1}\|_2^2 - \frac{1}{2}\|\boldsymbol{u}_{t+1}\|_2^2\right) \le 0.
\end{aligned}$$

Therefore,

$$\begin{aligned}
u_{t+1} - u_t &\le g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{1}{t+1}\ell(\boldsymbol{z}_{t+1}, D_t) \\
&\quad - \frac{1}{t+1}g_t'(D_t) \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{f_t'(D_t) - g_t'(D_t)}{t+1} \\
&\quad + \frac{\ell(\boldsymbol{z}_{t+1}, D_t) - f_t'(D_t)}{t+1} \\
&\le \frac{\ell(\boldsymbol{z}_{t+1}, D_t) - f_t'(D_t)}{t+1},
\end{aligned}$$

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 expectation on the above equation conditioned on $\mathcal{F}_t$:

$$\begin{aligned}\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t] &\le \frac{\mathbb{E}[\ell(\boldsymbol{z}_{t+1}, D_t) \mid \mathcal{F}_t] - f_t'(D_t)}{t+1} \\ &\le \frac{f(D_t) - f_t'(D_t)}{t+1} \\ &\le \frac{\|f - f_t'\|_\infty}{t+1}.\end{aligned}$$

From Proposition 11, we know

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

Thus,

$$\begin{aligned}\mathbb{E}[\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t]^+] &= \mathbb{E}[\max\{\mathbb{E}[u_{t+1} - u_t \mid \mathcal{F}_t], 0\}] \\ &\le \frac{c}{\sqrt{t}(t+1)},\end{aligned}$$

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

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

Thus, Lemma 5 applies. That is, $g_t(D_t)$ is a quasi-martingale and converges almost surely. Moreover,

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

□

F.4. Proof of Convergence of D_t

Proposition 13. *Let $\{D_t\}_{t=1}^{\infty}$ be the basis sequence produced by the Algorithm 1. Then,*

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

Proof. According the strong convexity of $g_t(D)$ (Assumption 3), we have,

$$g_t(D_{t+1}) - g_t(D_t) \ge \frac{\beta_0}{2} \|D_{t+1} - D_t\|_F^2 , \tag{F.18}$$

On the other hand,

$$\begin{aligned}& g_t(D_{t+1}) - g_t(D_t) \\ = {} & g_t(D_{t+1}) - g_{t+1}(D_{t+1}) + g_{t+1}(D_{t+1}) - g_{t+1}(D_t) \\ & + g_{t+1}(D_t) - g_t(D_t) \\ \le {} & g_t(D_{t+1}) - g_{t+1}(D_{t+1}) + g_{t+1}(D_t) - g_t(D_t) \\ \stackrel{\text{def}}{=} {} & G_t(D_{t+1}) - G_t(D_t).\end{aligned} \tag{F.19}$$

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

By a simple calculation, we obtain the gradient of $G_t(D)$:

$$\begin{aligned}& \nabla G_t(D) \\ = {} & \nabla g_t(D) - \nabla g_{t+1}(D) \\ = {} & \frac{1}{t}\Big[D\left(\lambda_1 A_t + \lambda_3 I_d\right) - \left(\lambda_1 B_t + \lambda_3 M_t\right)\Big] \\ & - \frac{1}{t+1}\Big[D(\lambda_1 A_{t+1} + \lambda_3 I_d) - (\lambda_1 B_{t+1} + \lambda_3 M_{t+1})\Big] \\ = {} & \frac{1}{t}\Bigg[D\left(\lambda_1 A_t + \lambda_3 I_d - \frac{\lambda_1 t}{t+1} A_{t+1} - \frac{\lambda_3 t}{t+1} I_d\right) \\ & + \frac{\lambda_1 t}{t+1} B_{t+1} - \lambda_1 B_t + \frac{\lambda_3 t}{t+1} M_{t+1} - \lambda_3 M_t\Bigg] \\ = {} & \frac{1}{t}\Bigg[D\left(\frac{\lambda_1}{t+1} A_{t+1} - \lambda_1 \boldsymbol{v}_{t+1}\boldsymbol{v}_{t+1}^\top + \frac{\lambda_3}{t+1} I_d\right) \\ & + \lambda_1(\boldsymbol{z}_{t+1} - \boldsymbol{e}_{t+1})\boldsymbol{v}_{t+1}^\top - \frac{\lambda_1}{t+1} B_{t+1} \\ & + \lambda_3 \boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top - \frac{\lambda_3}{t+1} M_{t+1}\Bigg]\end{aligned}$$

So the Frobenius norm of $\nabla G_t(D)$ follows immediately:

$$\begin{aligned}
&\|\nabla G_t(D)\|_F \\
\leq& \frac{1}{t}\Bigg[\|D\|_F \left(\lambda_1 \left\| \frac{A_{t+1}}{t+1} \right\|_F + \lambda_1 \left\| \boldsymbol{v}_{t+1}\boldsymbol{v}_{t+1}^\top \right\|_F \right. \\
&\left. + \frac{\lambda_3}{t+1} \|I_d\|_F \right) + \lambda_1 \left\| (\boldsymbol{z}_{t+1} - \boldsymbol{e}_{t+1})\boldsymbol{v}_{t+1}^\top \right\|_F \\
&+ \lambda_1 \left\| \frac{B_{t+1}}{t+1} \right\|_F + \lambda_3 \left\| \boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top \right\|_F \\
&+ \frac{\lambda_3}{t+1} \|M_{t+1}\|_F \Bigg] \\
=& \frac{1}{t}\Bigg[\|D\|_F \left(\lambda_1 \left\| \frac{A_{t+1}}{t+1} \right\|_F + \lambda_1 \left\| \boldsymbol{v}_{t+1}\boldsymbol{v}_{t+1}^\top \right\|_F \right) \\
&+ \lambda_1 \left\| (\boldsymbol{z}_{t+1} - \boldsymbol{e}_{t+1})\boldsymbol{v}_{t+1}^\top \right\|_F \\
&+ \lambda_1 \left\| \frac{B_{t+1}}{t+1} \right\|_F + \lambda_3 \left\| \boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top \right\|_F \Bigg] \\
&+ \frac{\lambda_3}{t(t+1)} \Big[\|I_d\|_F + \|M_{t+1}\|_F \Big].
\end{aligned}$$

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,

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

where α 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

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

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

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

□

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,

$$\begin{aligned}
b_t &= g_t(D_t) - f_t(D_t) \\
&= g_t'(D_t) - f_t'(D_t) + \Bigg[\frac{1}{t}\sum_{i=1}^{t} \frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2 \\
&\quad - \frac{1}{t^3}\sum_{i=1}^{t}\frac{1}{2}\left\| D_t^\top \left(\frac{1}{\lambda_3 t} I_p + \frac{1}{t} N_t \right)^{-1} \boldsymbol{z}_i \right\|_2^2 \\
&\quad - \frac{\lambda_3}{2t^3} \left\| \left(\frac{1}{t} I_p + \frac{\lambda_3}{t} N_t \right)^{-1} D_t \right\|_F^2 \Bigg] \\
&= g_t'(D_t) - f_t'(D_t) + q_t(D_t),
\end{aligned}$$

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

$$\begin{aligned}
\frac{b_t}{t+1} &= \frac{g_t'(D_t) - f_t'(D_t)}{t+1} + \frac{q_t(D_t)}{t+1} \\
&= g_{t+1}(D_{t+1}) - g_{t+1}(D_t) + \frac{\ell(\boldsymbol{z}_{t+1}, D_t) - f_t'(D_t)}{t+1} \\
&\quad + u_t - u_{t+1} \\
&\quad + \Bigg[\frac{q_t(D_t)}{t+1} + \frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 \\
&\quad + \frac{\lambda_3}{2(t+1)}\|D_t - M_{t+1}\|_F^2 - \frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 \\
&\quad - \frac{\lambda_3}{2t}\|D_t - M_t\|_F^2 \Bigg].
\end{aligned}$$

Note that we naturally have

$$\begin{aligned}
\frac{q_t(D_t)}{t+1} &\leq \frac{1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{\lambda_3}{2t(t+1)}\|D_t - M_t\|_F^2 \\
&\leq \frac{1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\|\boldsymbol{u}_i\|_2^2 + \frac{c}{2t(t+1)},
\end{aligned}$$

where the second inequality holds as D_t and M_t are both uniformly bounded (see Proposition 7).

Also, from (F.14), we know

$$\begin{aligned}
&\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2-\frac{1}{t}\sum_{i=1}^{t}\left\|\boldsymbol{u}_i\right\|_2^2\\
=&\frac{-1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2+\frac{1}{2(t+1)}\left\|\boldsymbol{u}_{t+1}\right\|_2^2.
\end{aligned}$$

And from (F.15)

$$\begin{aligned}
&\frac{\lambda_3}{2(t+1)}\left\|D_t-M_{t+1}\right\|_F^2-\frac{\lambda_3}{2t}\left\|D_t-M_t\right\|_F^2\\
\le&\frac{1}{t+1}\left(-\frac{\lambda_3}{2}\left\|\boldsymbol{z}_{t+1}\right\|_2^2\left\|\boldsymbol{u}_{t+1}\right\|_2^2-\left\|\boldsymbol{u}_{t+1}\right\|_2^2\right).
\end{aligned}$$

Thus,

$$\begin{aligned}
&\frac{q_t(D_t)}{t+1}+\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2\\
&+\frac{\lambda_3}{2(t+1)}\left\|D_t-M_{t+1}\right\|_F^2\\
&-\frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2-\frac{\lambda_3}{2t}\left\|D_t-M_t\right\|_F^2\\
\le&\frac{c}{2t(t+1)}+\frac{1}{2(t+1)}\left\|\boldsymbol{u}_{t+1}\right\|_2^2\\
&+\frac{1}{t+1}\left(-\frac{\lambda_3}{2}\left\|\boldsymbol{z}_{t+1}\right\|_2^2\left\|\boldsymbol{u}_{t+1}\right\|_2^2-\left\|\boldsymbol{u}_{t+1}\right\|_2^2\right)\\
=&\frac{c}{2t(t+1)}-\frac{1}{2(t+1)}\left\|\boldsymbol{u}_{t+1}\right\|_2^2\\
&-\frac{\lambda_3}{2(t+1)}\left\|\boldsymbol{z}_{t+1}\right\|_2^2\left\|\boldsymbol{u}_{t+1}\right\|_2^2\\
\le&\frac{c}{2t(t+1)}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\frac{b_t}{t+1}\\
\le&\, g_{t+1}(D_{t+1})-g_{t+1}(D_t)+\frac{\ell(\boldsymbol{z}_{t+1},D_t)-f_t'(D_t)}{t+1}\\
&+u_t-u_{t+1}+\frac{c}{2t(t+1)}\\
\le&\frac{\ell(\boldsymbol{z}_{t+1},D_t)-f_t'(D_t)}{t+1}+u_t-u_{t+1}+\frac{c}{2t(t+1)}.
\end{aligned}$$

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

$$\begin{aligned}
\frac{b_t}{t+1}\le&\frac{f(D_t)-f_t(D_t)}{t+1}+\mathbb{E}[u_t-u_{t+1}\mid\mathcal{F}_t]+\frac{c}{2t(t+1)}\\
\le&\frac{c_1}{\sqrt{t}(t+1)}+\left|\mathbb{E}[u_t-u_{t+1}\mid\mathcal{F}_t]\right|+\frac{c}{2t(t+1)},
\end{aligned}$$

where the second inequality holds by applying Proposition 11. Thus,

$$\begin{aligned}
&\sum_{t=1}^{\infty}\frac{b_t}{t+1}\\
\le&\sum_{t=1}^{\infty}\frac{c_1}{\sqrt{t}(t+1)}+\sum_{t=1}^{\infty}\left|\mathbb{E}[u_t-u_{t+1}\mid\mathcal{F}_t]\right|\\
&+\sum_{t=1}^{\infty}\frac{c}{2t(t+1)}\\
<&+\infty.
\end{aligned}$$

Here, the last inequality is derived by applying (F.16).

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

$$\begin{aligned}
&\left|b_{t+1}-b_t\right|\\
=&\left|g_{t+1}(D_{t+1})-f_{t+1}(D_{t+1})-g_t(D_t)+f_t(D_t)\right|\\
\le&\left|g_{t+1}(D_{t+1})-g_t(D_{t+1})\right|+\left|g_t(D_{t+1})-g_t(D_t)\right|\\
&+\left|f_{t+1}(D_{t+1})-f_t(D_{t+1})\right|+\left|f_t(D_{t+1})-f_t(D_t)\right|.
\end{aligned}\tag{F.20}$$

For the first term on the right hand side,

$$\begin{aligned}
&\left|g_{t+1}(D_{t+1})-g_t(D_{t+1})\right|\\
=&\Bigg|g_{t+1}'(D_{t+1})-g_t'(D_{t+1})+\frac{1}{t+1}\sum_{i=1}^{t+1}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2\\
&-\frac{1}{t}\sum_{i=1}^{t}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2+\frac{\lambda_3}{2(t+1)}\left\|D_{t+1}-M_{t+1}\right\|_F^2\\
&-\frac{\lambda_3}{2t}\left\|D_{t+1}-M_t\right\|_F^2\Bigg|\\
=&\Bigg|g_{t+1}'(D_{t+1})-g_t'(D_{t+1})-\frac{1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2\\
&-\frac{1}{2(t+1)}\left\|\boldsymbol{u}_{t+1}\right\|_2^2\\
&-\frac{\lambda_3}{2t(t+1)}\left\|D_{t+1}-M_t\right\|_F^2-\frac{\lambda_3}{2(t+1)}\left\|\boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right\|_F^2\Bigg|\\
\le&\left|g_{t+1}'(D_{t+1})-g_t'(D_{t+1})\right|+\frac{1}{t(t+1)}\sum_{i=1}^{t}\frac{1}{2}\left\|\boldsymbol{u}_i\right\|_2^2\\
&+\frac{1}{2(t+1)}\left\|\boldsymbol{u}_{t+1}\right\|_2^2+\frac{\lambda_3}{2t(t+1)}\left\|D_{t+1}-M_t\right\|_F^2\\
&+\frac{\lambda_3}{2(t+1)}\left\|\boldsymbol{z}_{t+1}\boldsymbol{u}_{t+1}^\top\right\|_F^2\\
\overset{\zeta_1}{\le}&\left|g_{t+1}'(D_{t+1})-g_t'(D_{t+1})\right|+\frac{c_1}{t+1}\\
=&\left|\frac{1}{t+1}\ell(\boldsymbol{z}_{t+1},D_{t+1})-\frac{1}{t+1}g_t'(D_{t+1})\right|+\frac{c_1}{t+1}\\
\overset{\zeta_2}{\le}&\frac{c_2}{t+1},
\end{aligned}$$

where c_1 and c_2 are some uniform constants. Note that ζ_1 holds because all the $\boldsymbol{u}_i$'s, D_{t+1}, M_t and $\boldsymbol{z}_{t+1}$ are uniformly bounded (see Proposition 7), and ζ_2 holds because $\ell(\boldsymbol{z}_{t+1}, D_{t+1})$ 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

$$\begin{aligned}
&|f_{t+1}(D_{t+1}) - f_t(D_{t+1})| \\
\leq & \left|f'_{t+1}(D_{t+1}) - f'_t(D_{t+1})\right| + \frac{c_3}{t+1} \\
= & \left|\frac{1}{t+1}\ell(\boldsymbol{z}_{t+1}, D_{t+1}) - \frac{1}{t+1}f'_t(D_{t+1})\right| + \frac{c_3}{t+1} \\
\overset{\zeta_3}{\leq} & \frac{c_4}{t+1},
\end{aligned}$$

where c_3 and c_4 are some uniform constants, and ζ_3 holds as $\ell(\boldsymbol{z}_{t+1}, D_{t+1})$ 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 κ_1, κ_2, such that

$$\begin{aligned}
|g_t(D_{t+1}) - g_t(D_t)| &\leq \kappa_1 \left\|D_{t+1} - D_t\right\|_F \overset{\zeta_4}{\leq} \frac{\kappa_3}{t+1}, \\
|f_t(D_{t+1}) - f_t(D_t)| &\leq \kappa_2 \left\|D_{t+1} - D_t\right\|_F \overset{\zeta_5}{\leq} \frac{\kappa_4}{t+1}.
\end{aligned}$$

Here, ζ_4 and ζ_5 are derived by applying Proposition 13 and κ_3 and κ_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 κ_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 $g_t(D_t)$. □

Theorem 15 (Convergence of $f(D_t)$)**.** *Let $f(D)$ 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 $\sqrt{t}(f(D_t) - f_t(D_t))$ is bounded, implying

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

□

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 $\nabla f(D) = \mathbb{E}_{\boldsymbol{z}}[\nabla_D \ell(\boldsymbol{z}, D)]$. Moreover, $\nabla f(D)$ is uniformly Lipschitz on $\mathcal{D}$.*

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

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

Let $\Lambda = \{j \mid \boldsymbol{e}_j' \neq 0\}$. According the first order optimality condition, we know that

$$\frac{\partial \tilde{\ell}(\boldsymbol{z}, D, \boldsymbol{v}, \boldsymbol{e})}{\partial \boldsymbol{e}} = 0,$$

which implies

$$\lambda_1(\boldsymbol{z} - D\boldsymbol{v} - \boldsymbol{e}) \in \lambda_2 \left\|\boldsymbol{e}\right\|_1.$$

Hence,

$$|(\boldsymbol{z} - D\boldsymbol{v} - \boldsymbol{e})_j| = \frac{\lambda_2}{\lambda_1},\ \forall j \in \Lambda.$$

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

Let us denote $H = [D\ I_p]$, $\boldsymbol{r} = [\boldsymbol{v}^\top\ \boldsymbol{e}^\top]^\top$ and define the function

$$\begin{aligned}
\tilde{\ell}(\boldsymbol{z}, H, \boldsymbol{r}_\Lambda) = &\frac{\lambda_1}{2}\left\|\boldsymbol{z} - H_\Lambda \boldsymbol{r}_\Lambda\right\|_2^2 + \frac{1}{2}\left\|[I\ 0]\boldsymbol{r}_\Lambda\right\|_2^2 \\
&+ \lambda_2 \left\|[0\ I]\boldsymbol{r}_\Lambda\right\|_1
\end{aligned}$$

Since $\tilde{\ell}(\boldsymbol{z}, D_\Lambda, \cdot)$ is strongly convex, there exists a uniform constant κ_1, such that for all $\boldsymbol{r}''_\Lambda$,

$$\begin{aligned}
&\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}''_\Lambda) - \tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}'_\Lambda) \\
\geq &\kappa_1 \left\|\boldsymbol{r}''_\Lambda - \boldsymbol{r}'_\Lambda\right\|_2^2 \\
= &\kappa_1 \left(\left\|\boldsymbol{v}'' - \boldsymbol{v}'\right\|_2^2 + \left\|\boldsymbol{e}''_\Lambda - \boldsymbol{e}'_\Lambda\right\|_2^2\right).
\end{aligned} \tag{F.21}$$

On the other hand,

$$\begin{aligned}
&\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}''_\Lambda) - \tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}'_\Lambda) \\
= &\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}''_\Lambda) - \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}''_\Lambda) \\
&+ \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}''_\Lambda) - \tilde{\ell}(\boldsymbol{z}', D'_\Lambda, \boldsymbol{r}'_\Lambda) \\
\leq &\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}''_\Lambda) - \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}''_\Lambda) \\
&+ \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}'_\Lambda) - \tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}'_\Lambda),
\end{aligned} \tag{F.22}$$

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

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

$$
\begin{aligned}
&\nabla_{\boldsymbol{r}} \left(\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}_\Lambda) - \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}_\Lambda) \right) \\
&= \lambda_1 \Big[H'^\top_\Lambda (H'_\Lambda - H''_\Lambda) + (H'_\Lambda - H''_\Lambda)^\top H''_\Lambda \\
&\quad + H'^\top_\Lambda (\boldsymbol{z}'' - \boldsymbol{z}') + (H''_\Lambda - H'_\Lambda)^\top \boldsymbol{z}'' \Big].
\end{aligned}
$$

Note that $\|H'_\Lambda\|_F$, $\|H''_\Lambda\|_F$ and $\boldsymbol{z}''$ are all uniformly bounded. Hence, there exists uniform constants c_1 and c_2, such that

$$
\begin{aligned}
&\left\| \nabla_{\boldsymbol{r}} \left(\tilde{\ell}(\boldsymbol{z}', H'_\Lambda, \boldsymbol{r}_\Lambda) - \tilde{\ell}(\boldsymbol{z}'', H''_\Lambda, \boldsymbol{r}_\Lambda) \right) \right\|_2 \\
&\le c_1 \|H'_\Lambda - H''_\Lambda\|_F + c_2 \|\boldsymbol{z}' - \boldsymbol{z}''\|_2 ,
\end{aligned}
$$

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

$$
\kappa_1 \|\boldsymbol{r}''_\Lambda - \boldsymbol{r}'_\Lambda\|_2^2 \le c(H'_\Lambda, H''_\Lambda, \boldsymbol{z}', \boldsymbol{z}'') \|\boldsymbol{r}''_\Lambda - \boldsymbol{r}'_\Lambda\|_2 .
$$

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

$$
\begin{aligned}
&\nabla f(D') - \nabla f(D'') \\
&= \mathbb{E}_{\boldsymbol{z}} \Big[(H'\boldsymbol{r}' - \boldsymbol{z})\boldsymbol{v}'^\top - (H''\boldsymbol{r}'' - \boldsymbol{z})\boldsymbol{v}''^\top \Big] \\
&= \mathbb{E}_{\boldsymbol{z}} \Big[H'\boldsymbol{r}'(\boldsymbol{v}' - \boldsymbol{v}'')^\top + (H' - H'')\boldsymbol{r}'\boldsymbol{v}''^\top \\
&\quad + H''(\boldsymbol{r}' - \boldsymbol{r}'')\boldsymbol{v}''^\top + \boldsymbol{z}(\boldsymbol{v}'' - \boldsymbol{v}')^\top \Big].
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&\|\nabla f(D') - \nabla f(D'')\|_F \\
&\overset{\zeta_1}{\le} \mathbb{E}_{\boldsymbol{z}} \Big[\|H'\boldsymbol{r}'\|_2 \|\boldsymbol{v}' - \boldsymbol{v}''\|_2 + \|H' - H''\|_F \left\|\boldsymbol{r}'\boldsymbol{v}''^\top\right\|_F \\
&\quad + \|H''\|_F \|\boldsymbol{r}' - \boldsymbol{r}''\|_2 \|\boldsymbol{v}''\|_2 + \|\boldsymbol{z}\|_2 \|\boldsymbol{v}' - \boldsymbol{v}''\|_2 \Big] \\
&\overset{\zeta_2}{\le} \mathbb{E}_{\boldsymbol{z}} \Big[(\gamma_1 + \gamma_2 \|\boldsymbol{z}\|_2) \|H' - H''\|_F \Big] \\
&\overset{\zeta_3}{\le} \gamma_0 \|D' - D''\|_F ,
\end{aligned}
$$

where γ_0, γ_1 and γ_2 are all uniform constants. Here, ζ_1 holds because for any function $s(\boldsymbol{z})$, we have $\|\mathbb{E}_{\boldsymbol{z}}[s(\boldsymbol{z})]\|_F \le \mathbb{E}_{\boldsymbol{z}}[\|s(\boldsymbol{z})\|_F]$. ζ_2 is derived by using the result that $\boldsymbol{r}(\boldsymbol{z}, H)$ and $\boldsymbol{v}(\boldsymbol{z}, H)$ are both Lipschitz and H', H'', $\boldsymbol{r}'$, $\boldsymbol{r}''$, $\boldsymbol{v}'$ and $\boldsymbol{v}''$ are all uniformly bounded. ζ_3 holds because $\boldsymbol{z}$ is uniformly bounded and actually $\|H' - H''\|_F = \|D' - D''\|_F$. 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 $\{\frac{1}{t}A_t\}$ and $\{\frac{1}{t}B_t\}$ that converge to A_∞ and B_∞ respectively. Then D_t will converge to D_∞. 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

$$
g_t(D_t + h_k W) \ge f_t(D_t + h_k W).
$$

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

$$
g_\infty(D_\infty + h_k W) \ge f(D_\infty + h_k W).
$$

Note that the Lipschitz of ∇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,

$$
\begin{aligned}
&\mathrm{Tr}(h_k W^\top \nabla g_\infty(D_\infty)) + o(h_k W) \\
\ge\ &\mathrm{Tr}(h_k W^\top \nabla f(D_\infty)) + o(h_k W)
\end{aligned}
$$

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

$$
\begin{aligned}
&\mathrm{Tr}\left(\frac{1}{\|W\|_F} W^\top \nabla g_\infty(D_\infty) \right) + \frac{o(h_k W)}{h_k \|W\|_F} \\
\ge\ &\mathrm{Tr}\left(\frac{1}{\|W\|_F} W^\top \nabla f(D_\infty) \right) + \frac{o(h_k W)}{h_k \|W\|_F}.
\end{aligned}
$$

Now let k go to infinity,

$$
\mathrm{Tr}\left(\frac{1}{\|W\|_F} W^\top \nabla g_\infty(D_\infty) \right) \ge \mathrm{Tr}\left(\frac{1}{\|W\|_F} W^\top \nabla f(D_\infty) \right).
$$

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_∞ is the minimizer of $g_\infty(D)$, we have

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

□

References

Avron, Haim, Kale, Satyen, Kasiviswanathan, Shiva Prasad, and Sindhwani, Vikas. Efficient and practical stochastic subgradient descent for nuclear norm regularization. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.

Bertsekas, Dimitri P. *Nonlinear programming*. Athena Scientific, 1999.

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

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

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

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

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

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

Defazio, Aaron, Bach, Francis R., and Lacoste-Julien, Simon. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Neural Information Processing Systems*, pp. 1646–1654, 2014.

Elhamifar, Ehsan and Vidal, René. Sparse subspace clustering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2790–2797, 2009.

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

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

Feng, Jiashi, Xu, Huan, and Yan, Shuicheng. Online robust PCA via stochastic optimization. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 404–412, 2013.

Guo, Han, Qiu, Chenlu, and Vaswani, Namrata. An online algorithm for separating sparse and low-dimensional signal sequences from their sum. *IEEE Trans. Signal Processing*, 62(16):4284–4297, 2014.

Hsieh, Cho-Jui and Olsen, Peder A. Nuclear norm minimization via active subspace selection. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 575–583, 2014.

Jaggi, Martin and Sulovský, Marek. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 471–478, 2010.

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

Liu, Guangcan and Li, Ping. Recovery of coherent data via low-rank dictionary pursuit. In *Proceedings of the 28th Annual Conference on Neural Information Processing Systems*, pp. 1206–1214, 2014.

Liu, Guangcan, Lin, Zhouchen, Yan, Shuicheng, Sun, Ju, Yu, Yong, and Ma, Yi. Robust recovery of subspace structures by low-rank representation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 35(1):171–184, 2013.

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

Mardani, Morteza, Mateos, Gonzalo, and Giannakis, Georgios B. Subspace learning and imputation for streaming big data matrices and tensors. *IEEE Trans. Signal Processing*, 63(10):2663–2677, 2015.

Ng, Andrew Y., Jordan, Michael I., and Weiss, Yair. On spectral clustering: Analysis and an algorithm. In *Proceedings of the 15th Annual Conference on Neural Information Processing Systems*, pp. 849–856, 2001.

Qiu, Chenlu, Vaswani, Namrata, Lois, Brian, and Hogben, Leslie. Recursive robust PCA or recursive sparse recovery in large but structured noise. *IEEE Trans. Information Theory*, 60(8):5007–5039, 2014.

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

Richtárik, Peter and Takác, Martin. Iteration complexity of randomized block-coordinate descent methods for minimizing a composite function. *Mathematical Programming*, 144(1-2):1–38, 2014.

Shen, Jie and Li, Ping. Learning structured low-rank representation via matrix factorization. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, pp. 500–509, 2016.

Soltanolkotabi, Mahdi and Candès, Emmanuel J. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40:2195–2238, 2012.

Soltanolkotabi, Mahdi, Elhamifar, Ehsan, and Candès, Emmanuel J. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

van der Vaart, A.W. *Asymptotic statistics*. Cambridge University Press, 2000.

Vidal, René. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2010.

Wang, Yu-Xiang, Xu, Huan, and Leng, Chenlei. Provable subspace clustering: When LRR meets SSC. In *Proceedings of 27th Annual Conference on Neural Information Processing Systems*, pp. 64–72, 2013.

Xiao, Lin and Zhang, Tong. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

Xu, Huan, Caramanis, Constantine, and Mannor, Shie. Principal component analysis with contaminated data: The high dimensional case. In *Proceedings of the 23rd Conference on Learning Theory*, pp. 490–502, 2010.