6. Appendix

6.1. Proof sketch for Theorem 2

This theorem is proved in (Mazumder et al., 2010) by considering the auxilliary function

$$\begin{aligned} Q(X,Y) &= \frac{1}{2}\|\Pi_\Omega(A)+\Pi_\Omega^\perp(Y)-X\|_F^2+\lambda\|X\|_* \\ &= F(X)+\frac{1}{2}\|\Pi_\Omega^\perp(Y-X)\|_F^2, \end{aligned}$$

for which $Q(X,Y) \geq F(X)$ and $Q(X,X) = F(X)$. We can minimize the auxiliary function by noting that the minimum with respect to Y for fixed X is $Y = X$ and for fixed Y the minimum with respect to X is $X = S_\lambda(\Pi_\Omega(A)+\Pi_\Omega^\perp(Y))$. Alternating the minimization gives the iteration in the theorem. This algorithm is known as Soft-Impute.

6.2. Proof sketch for Theorem 3

For the regression problem we can form a different auxilliary function. If $cI \succ A^\top A = f''(X)$ then $-\frac{1}{2}\|AX-AY\|_F^2+\frac{c}{2}\|X-Y\|_F^2 \geq 0$ for all X, Y and the auxilliary function

$$\begin{aligned} Q(X,Y) &= \frac{1}{2}\|AX-B\|_F^2-\frac{1}{2}\|A(X-Y)\|_F^2 \\ &\quad +\frac{c}{2}\|X-Y\|_F^2+\lambda\|X\|_* \\ &= \frac{c}{2}\|X-Y-\frac{1}{c}(A^\top B-A^\top AY)\|_F^2 \\ &\quad +\lambda\|X\|_*+\text{const} \\ &= c(\frac{1}{2}\|X^\top-(Y-\frac{1}{c}f'(Y))\|_F^2+\frac{\lambda}{c}\|X\|_*) \\ &\quad +\text{const} \end{aligned}$$

satisfy $Q(X,Y) \geq F(X)$ and $Q(X,X) = F(X)$. For fixed Y_k the minimum over X is $X_{k+1} = S_{\lambda/c}(Y_k - \frac{1}{c}f'(Y_k))$ and for fixed X_k the minimum over Y is $Y_k = X_k$. This auxilliary function is constructed completely analogously to the ℓ_1 case, for which global convergence is formally proved in (Daubechies et al., 2004).

6.3. Proof of Theorem 4

Proof. If $\mathbf{u} \in U_A^\perp$, then (1) $U^\top\mathbf{u} = 0$, which implies $\mathbf{u}^\top X = 0$; (2) $U_G^\top\mathbf{u} = 0$, which implies $|\mathbf{u}^\top(X-\nabla f(X))\mathbf{v}| < \lambda$ for any $\mathbf{v}$ (by the definition of soft-thresholding operator $\mathcal{S}$). Combining (1) and (2) we have $\mathbf{u}\mathbf{v}^\top \in \mathcal{F}$ for all $\mathbf{v}$ if $\mathbf{u} \in U_A^\perp$. By the same argument we can prove $\mathbf{u}\mathbf{v}^\top \in \mathcal{F}$ for all $\mathbf{u}$ if $\mathbf{v} \in V_A^\perp$. □

6.4. Proof of Theorem 6

Proof. Since S is positive definite it has an eigenvalue decomposition $S = P\Sigma P^\top$ with $\Sigma \succ 0$ a diagonal matrix. Therefore the SVD of X can be written $X = (UP)\Sigma(VP)^\top$ and the sub-differential is

$$\partial\|X\|_* = \{UV^\top+W : U^\top W = 0, WV = 0, \|W\|_2 \leq 1\},$$

independent of S since $(UP)(VP)^\top = UV^\top$. □

6.5. Proof of Theorem 7

Proof. Assume $X^* = U^*\Sigma^*V^*$ is the reduced SVD of X^*. Since X^* is the global optimum,

$$\begin{aligned} X^* &= S_\lambda(X^*-\nabla f(X^*)) \\ &= \bar{U}^*(\bar{\Sigma}^*-\lambda I)_+(\bar{V}^*)^\top. \end{aligned} \tag{15}$$

If there are k singular values in $\bar{\Sigma}^*$ larger than λ, then it is clear that the first k columns of $\bar{U}^*$ is U^*, and the first k columns of $\bar{V}^*$ is V^*. By our assumption, $\Sigma_{ii} \neq \lambda$ for all i, so we can assume $\Sigma_{kk} > \lambda$ and $\Sigma_{k+1,k+1} < \lambda-\epsilon$ with some $\epsilon > 0$.

We consider the set

$$\mathcal{Z} \equiv \{(\mathbf{u},\mathbf{v}) \mid \mathbf{u} \in (U^*)^\perp \text{ or } \mathbf{v} \in (V^*)^\perp\}.$$

For $(\mathbf{u},\mathbf{v}) \in \mathcal{Z}$, $\mathbf{u}^\top X^*\mathbf{v} = 0$, so

$$|\mathbf{u}^\top(X^*-\nabla f(X^*))\mathbf{v}| = |\mathbf{u}^\top\nabla f(X^*)\mathbf{v}| < \lambda-\epsilon.$$

Since the sequence X_t generated by Algorithm 1 converges to the global optimum X^*, there exists a T such that

$$\|\nabla f(X_t)-\nabla f(X^*)\| < \epsilon \tag{16}$$

and

$$|\mathbf{u}^\top\nabla f(X_t)\mathbf{v}| < \lambda \tag{17}$$

for all $t > T$ and any $(u,v) \in \mathcal{Z}$. Now for any $(u,v) \in \mathcal{Z}$ we consider two cases:

1. If $\mathbf{u}^\top X_{t-1}\mathbf{v} \neq 0$, then $\mathbf{u} \in (U_A)_{t-1}$ and $\mathbf{v} \in (V_A)_{t-1}$. Since we exactly solve the sub-problem (7) and we already know $|\mathbf{u}^\top\nabla f(X_t)\mathbf{v}| < \lambda$, the optimality condition of (7) implies $\mathbf{u}^\top X_t\mathbf{v} = 0$.
2. If $\mathbf{u}^\top X_{t-1}\mathbf{v} = 0$, then combined with (17) we know $\mathbf{u}, \mathbf{v}$ are not in the active subspace, so $\mathbf{u}^\top X_t\mathbf{v} = 0$.

Therefore, once $t > T$, for any $\mathbf{u} \in (U^*)^\perp$ or $\mathbf{v} \in (V^*)^\perp$, $\mathbf{u}^\top X_t\mathbf{v}$ will be zero and will never be selected in $(U_A)_t, (V_A)_t$. This implies that span$((U_A)_t) \subseteq$ span(U^*) and span$((V_A)_t) \subseteq$ span(V^*).

Next we prove the equality part. For all $\mathbf{u}, \mathbf{v}$ such that $\mathbf{u}^\top X^*\mathbf{v} \neq 0$, there exists a T such that $\mathbf{u}^\top(X_t)\mathbf{v} \neq 0$ for all $t > T$ (since the smallest eigenvalue > 0). Therefore, all such $\mathbf{u}, \mathbf{v}$ will belong to $(U_A)_t, (V_A)_t$ after $t > T$. Combined with the previous argument, we have span$((U_A)_t) =$ span(U^*) and span$((V_A)_t) =$ span(V^*) after $t > T$. □

6.6. Proof of Theorem 9

Proof. We first introduce an important property of the power method, which will be useful for proving the theorem.

The power method (subspace iteration) described in Algorithm 2 has a linear convergence rate: assume U, V are the top-k singular vectors of A, σ_k, σ_{k+1} are the k-th and $(k+1)$-st singular values, and the approximate SVD given by Algorithm 2 with R as initial and with T^{max} steps. If the initial matrix R satisfies the condition that $V^\top R$ is non-singular, then

$$\|\hat{U}\hat{U}^\top - UU^\top\| \le (\frac{\sigma_{k+1}}{\sigma_k})^{T^{max}} \|U_R U_R^\top - UU^\top\|,$$
$$\|\hat{V}\hat{V}^\top - VV^\top\| \le (\frac{\sigma_{k+1}}{\sigma_k})^{T^{max}} \|V_R V_R^\top - VV^\top\|. \quad (18)$$

where U_R is the orthogonal subspace of R, V_R is the orthogonal subspace of AR, and $\hat{U}$ is the subspace after one power iteration. This property is shown in Theorem 7.2 in (Arbenz, 2010).

Now we prove that the sequence X_t generated by Algorithm 1 converges to the global optimum. For convenience, we define $P(X) := \mathcal{S}_\lambda(X - \nabla f(X))$, and $\tilde{P}(X)$ to be the *computed* value (by the power method with one iteration) of $P(X)$. The reduced SVD of $\mathcal{S}_\lambda(X - \nabla f(X))$ is denoted by $U_G(X)\Sigma_G(X)(V_G(X))^\top$, and the computed subspace vectors is $\tilde{U}_G(X), \tilde{V}_G(X)$. We use $\tilde{U}_G(X_t)$ to denote the computed value at the t-th iteration, and $U_G(X_t)$ to denote the true subspace vectors at the t-th iteration.

Since Algorithm 1 ensures that the objective function value decreases at each iteration, the sequence $\{X_t\}$ is in a compact set. Therefore, there exists a subsequence of X_{s_t} converges to a limit point $\bar{X}$. For convenience we denote s_t by t in the following. We want to prove $\bar{X}$ is the global optimum by contradiction, so we first assume $\bar{X} \neq X^*$, so $P(\bar{X}) \neq \bar{X}$.

First we want to show $\tilde{U}(X_t), \tilde{V}(X_t)$ converges to $U_G(\bar{X}), V_G(\bar{X})$ (the computed subspace converges to the true subspace). Assume $\tilde{U}_G(X_t), \tilde{V}_G(X_t)$ converges to $\tilde{U}, \tilde{V}$, then what we want to show is that span$(\tilde{U})$ = span$(U_G(\bar{X}))$ and span$(\tilde{V})$ = span$(V_G(\bar{X}))$. Since $\{X_t\}$ converges to $\bar{X}$ and $X - \nabla f(X)$ is a continuous function, for any $\epsilon > 0$ there exists a T_1 such that $\forall t > T_1$,

$$\|(X_t - \nabla f(X_t)) - (\bar{X} - \nabla f(\bar{X}))\| \le \epsilon. \quad (19)$$

By perturbation theory (Li, 1998), for any matrix A and a small perturbation Δ, we have

$$\max(\|U(A)U(A)^\top - U(A+\Delta)U(A+\Delta)^\top\|,$$
$$\|V(A)V(A)^\top - V(A+\Delta)V(A+\Delta)^\top\|) \le \|\Delta\|/\delta,$$

where δ is the singular-gap between $\sigma_k(A)$ and $\sigma_{k+1}(A)$, and $U(A), V(A)$ are the top-k singular vectors of A. Now we consider $A = P(\bar{X}), \Delta = P(X_t) - P(\bar{X})$, then we have

$$\max(\|U_G(X_t)U_G(X_t)^\top - U_G(\bar{X})U_G(\bar{X})^\top\|,$$
$$\|V_G(X_t)V_G(X_t)^\top - V_G(\bar{X})V_G(\bar{X})\|) \le \|P(X_t) - P(\bar{X})\|/\delta,$$

Combining with (19) we get

$$\|U_G(X_t)U_G(X_t)^\top - U_G(\bar{X})U_G(\bar{X})^\top\| \le \frac{\epsilon}{\delta} \ \ \forall t > T_1. \quad (20)$$

Now assume t is large enough so that

$$\|\tilde{U}\tilde{U}^\top - \tilde{U}_G(X_{t-1})\tilde{U}_G(X_{t-1}^\top)\| < \epsilon_1, \quad (21)$$

so we have

$$\begin{aligned}
&\|\tilde{U}_G(X_t)\tilde{U}_G(X_t)^\top - U_G(\bar{X})U_G(\bar{X})^\top\| \\
&\le \|\tilde{U}_G(X_t)\tilde{U}_G(X_t)^\top - U_G(X_t)U_G(X_t)^\top\| + \frac{\epsilon}{\delta} \text{ (by (20))} \\
&\le (\frac{\sigma_{k+1}}{\sigma_k})\|\tilde{U}_G(X_{t-1})\tilde{U}_G(X_{t-1})^\top - U_G(X_t)U_G(X_t)^\top\| + \frac{\epsilon}{\delta} \text{ (by (18))} \\
&\le (\frac{\sigma_{k+1}}{\sigma_k})\|\tilde{U}\tilde{U}^\top - U_G(X_t)U_G(X_t)^\top\| + \frac{\epsilon}{\delta} + \epsilon_1. \text{ (by (21))} \\
&\le (\frac{\sigma_{k+1}}{\sigma_k})\|\tilde{U}\tilde{U}^\top - U_G(\bar{X})U_G(\bar{X})^\top\| + 2\frac{\epsilon}{\delta} + \epsilon_1 \text{ (by (20))}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\|\tilde{U}_G(X_t)\tilde{U}_G(X_t)^\top - \tilde{U}\tilde{U}^\top\| \\
&\ge \|U_G(\bar{X})U_G(\bar{X})^\top - \tilde{U}\tilde{U}^\top\| \\
&- \|\tilde{U}_G(X_t)\tilde{U}_G(X_t) - U_G(\bar{X})U_G(\bar{X})^\top\| \\
&\ge (1 - \frac{\sigma_{k+1}}{\sigma_k})\|\tilde{U}\tilde{U}^\top - U_G(\bar{X})U_G(\bar{X})^\top\| - 2\frac{\epsilon}{\delta} - \epsilon_1.
\end{aligned}$$

Taking $t \to \infty$ on both side and $\epsilon, \epsilon_1 \to 0$ we have

$$0 \ge (1 - \sigma_{k+1}/\sigma_k)\|\bar{U}\bar{U}^\top - U_G(\bar{X})U_G(\bar{X})^\top\|.$$

So span$(U_G(\bar{X}))$ = span$(\tilde{U})$. Using the same derivations on the right singular vectors V, we can get span$(V_G(\bar{X}))$ = span$(\tilde{V})$.

The above argument shows that $\tilde{P}(X_t) \to P(\bar{X})$. If $\bar{X}$ is not a global optimum, then $P(\bar{X}) \neq \bar{X}$. Since all the fixed points are global optimum, by a typical convergence property for the fixed-point operation we can show that $\bar{X}$ is a global optimum.

Next, we prove the asymptotic convergence rate. By Theorem 7, we know after finite steps T_1, $U_A(X_t) = U^*, V_A(X_t) = V^*$. Moreover, $\sigma_k(X_t - \nabla f(X_t))$ converges to $\sigma_k(X^* - \nabla f(X^*))$ and $\sigma_{k+1}(X_t - \nabla f(X_t))$ converges to $\sigma_{k+1}(X^* - \nabla f(X^*))$, so there exists a T_2 such that for all $t > T_2$,

$$\frac{\sigma_{k+1}(X_t - \nabla f(X_t))}{\sigma_k(X_t - \nabla f(X_t))} \le \frac{\lambda - \epsilon/2}{\lambda}. \quad (22)$$

Assume $\bar{T} = \max(T_1, T_2)$. Since for each iteration we run one power iteration on $X_t - \nabla f(X_t)$ and the gap of the k-th and $(k+1)$-st singular values are guaranteed in (22), from (18) we can bound the error between subspaces $(U_A)_t$ and U^* by

$$\begin{aligned}&\|U_A(X_t)U_A(X_t)^\top - U^*(U^*)^\top\| \\ &\leq (1 - \frac{\epsilon}{2\lambda})\|U_A(X_{t-1})U_A(X_{t-1})^\top - U^*(U^*)^\top\|\end{aligned}$$

when $t > \bar{T}$. Therefore

$$\begin{aligned}&\|U_A(X_{\bar{T}+t})U_A(X_{\bar{T}+t})^\top - U^*(U^*)^\top\| \\ &\leq (1 - \frac{\epsilon}{2\lambda})^t\|U_A(X_{\bar{T}})U_A(X_{\bar{T}})^\top - U^*(U^*)^\top\|. \quad (23)\end{aligned}$$

At the t-th iteration, it is clear that $\bar{S}_t = U_A(X_t)^\top X^* V_A(X_t)$ is a feasible solution for the sub-problem (8). Let $\bar{X}_t = U_A(X_t)\bar{S}_t V_A(X_t)^\top$ then $F(X_t) \geq F(\bar{X}_t)$ (because X_t is the minimizer of (8)). Also, $X^* = U^*(U^*)^\top X^* = X^* V^*(V^*)^\top$, so

$$\begin{aligned}&\|\bar{X}_t - X^*\| \\ &\leq \|U_A(X_t)U_A(X_t)^\top X^* V_A(X_t)V_A(X_t)^\top - U_A(X_t)U_A(X_t)^\top X^*\| \\ &\quad + \|U_A(X_t)U_A(X_t)^\top X^* - X^*\| \\ &= \|U_A(X_t)U_A(X_t)^\top X^* \left(V^*(V^*)^\top - V_A(X_t)V_A(X_t)^\top\right)\| \\ &\quad + \|(U^*(U^*)^\top - U_A(X_t)U_A(X_t)^\top)X^*\| \\ &\leq (\|U^*(U^*)^\top - U_A(X_t)U_A(X_t)^\top\| + \\ &\quad \|V^*(V^*)^\top - V_A(X_t)V_A(X_t)^\top\|)\|X^*\|.\end{aligned}$$

Next we relate this quantity with the objective function value $F(X_t)$. From Lemma 3.1 in (Ji & Ye, 2009),

$$F(X) - F(X^*) \leq L\|X - X^*\|_F^2,$$

where L is the Lipschitz constant for $\nabla f(X)$. Substituting $\bar{X}_t$ into the above inequality we get

$$\begin{aligned}F(X_t) - F(X^*) &\leq F(\bar{X}_t) - F(X^*) \\ &\leq LR(\|U^*(U^*)^\top - U_A(X_t)U_A(X_t)^\top\| \\ &\quad + \|V^*(V^*)^\top - V_A(X_t)V_A(X_t)^\top\|),\end{aligned}$$

where $R = \|X^*\|$ is a constant. Applying (23) we can get

$$F(X_t) - F(X^*) \leq LR(1 - \frac{\epsilon}{2\lambda})^{t-\bar{T}}$$

when $t > \bar{T}$. Therefore our algorithm has an asymptotically linear convergence rate.

□

6.7. Implementation Details for the comparison

We discuss the implementation detail for other algorithms in our comparison. The code for Soft-Impute is downloaded from `http://statweb.stanford.edu/~rahulm/SoftImpute/`. In their code, the top-k singular vectors is computed by Lanczos algorithm. We use the same JSH and SSGD implementation as in (Avron et al., 2012), where the largest singular value is computed by the `SVDS` function in MATLAB and the parameters are tuned by the authors. More specifically, $\delta = 0.04$ for ml100k, $\delta = 0.015$ for ml10m and netflix, and $\nu = 0.005$ for all datasets. We implement LiftedCD by ourselves and compute the largest singular value by the power method. For MMBS the code is downloaded from `http://www.montefiore.ulg.ac.be/~mishra/softwares/traceNorm.html`, and the GCG code is downloaded from `http://users.cecs.anu.edu.au/~xzhang/GCG/`.