8 PROOFS OF LEMMATA

8.1 Proof of Lemma 4.1

Due to the decomposability of (7), we observe $\forall X$:

$$\frac{\partial h(X,\tau)}{\partial X_{ij}} = \frac{2X_{ij}}{\tau} \cdot \left(\left(\frac{X_{ij}}{\tau} \right)^2 + 1 \right)^{-1/2} = \frac{X_{ij}}{\tau} \cdot \frac{2}{\sqrt{\left(\frac{X_{ij}}{\tau} \right)^2 + 1}}$$

Thus, in compact form, $\nabla h(X,\tau) = \frac{1}{\tau} X \odot S$, where S is defined in the lemma.

Regarding the Hessian information, first observe that $\frac{\partial^2 h(X,\tau)}{\partial X_{ij} \partial X_{lq}} = \frac{\partial \left(\frac{X_{ij}}{\tau} \cdot \frac{2}{\sqrt{\left(\frac{X_{ij}}{\tau} \right)^2 + 1}} \right)}{\partial X_{lq}} = 0$, for indices $(i,j) \neq (l,q)$. This means that the off-diagonals of $\nabla^2 h(X,\tau)$ are zero. For the case where $(i,j) = (l,q)$, we have:

$$\frac{\partial^2 h(X,\tau)}{\partial X_{ij}^2} = \frac{\partial \left(\frac{X_{ij}}{\tau} \cdot \frac{2}{\sqrt{\left(\frac{X_{ij}}{\tau} \right)^2 + 1}} \right)}{\partial X_{ij}} = \frac{2}{\tau} \cdot \frac{\sqrt{(X_{ij}/\tau)^2 + 1} - \frac{X_{ij}^2}{\tau^2} \cdot \left((X_{ij}/\tau)^2 + 1 \right)^{-1/2}}{\left(\frac{X_{ij}}{\tau} \right)^2 + 1}$$

$$= \frac{2}{\tau} \cdot \frac{(X_{ij}/\tau)^2 + 1 - (X_{ij}/\tau)^2}{\left(\left(\frac{X_{ij}}{\tau} \right)^2 + 1 \right)^{3/2}} = \frac{1}{\tau} \cdot \frac{2}{\left(\left(\frac{X_{ij}}{\tau} \right)^2 + 1 \right)^{3/2}}$$

Then, $\nabla^2 h(X,\tau) = \frac{1}{\tau} I \odot Q$, where Q is defined in the lemma.

8.2 Proof of Lemma 4.2

The first part of the lemma is easily deduced from Lemma 4.1. Observe that $0 \preceq \nabla^2 h(X,\tau) \preceq \frac{2}{\tau} I, \ \forall X$; that is h function is convex with Lipschitz constant $\frac{2}{\tau}$. Moreover, by combining h with any strongly convex function $\psi(\cdot)$, say $\psi(X) := \frac{\lambda}{2}|X|_2^2$, we easily observe that the composite form $h(X,\tau) + \psi(X)$ satisfies $\lambda I \preceq \nabla^2 h(X,\tau) + \nabla^2 \psi(X) \preceq \left(\frac{2}{\tau} + \lambda \right) I$; *i.e.*, the composite form is also strongly convex.

The last part of the lemma is true because

$$|X|_1 \geq h(X,\tau) = \sum_{i=1}^{m} \sum_{j=1}^{n} h(X_{ij},\tau) = \tau \cdot \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\sqrt{\left(\frac{X_{ij}}{\tau} \right)^2 + 1} - 1 \right) = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\sqrt{X_{ij}^2 + \tau^2} - \tau \right)$$

$$\geq \sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}| - mn\tau = |X|_1 - mn\tau.$$

8.3 Proof of Lemma 4.3

The proof is elementary as in Lemma 4.1 and we state it for completeness. First, observe that (9) can be re-written as follows:

$$\sigma(X,\tau) = \tau \cdot \log \left(\frac{\mathrm{Tr}(\mathbb{1} \cdot P)}{2mn} \right)$$

Observe that calculating gradients with respect to X_{ij}, the denominator $2mn$ plays no role. Following similar motions, we compute partial derivatives as:

$$\frac{\partial \sigma(X,\tau)}{\partial X_{ij}} = \tau \cdot \frac{1}{\mathrm{Tr}(\mathbb{1} \cdot P)} \cdot \frac{\partial \left(e^{X_{ij}/\tau} + e^{-X_{ij}/\tau} \right)}{\partial X_{ij}} = \frac{1}{\mathrm{Tr}(\mathbb{1} \cdot P)} \cdot \left(e^{X_{ij}/\tau} - e^{-X_{ij}/\tau} \right)$$

Gathering all the partial derivatives in a matrix, we get the reported result.

Computing second-order partial derivatives for $\sigma(X,\tau)$, we distinct the cases of diagonal and off-diagonal elements. For the former, we have:

$$\frac{\partial^2\sigma(X,\tau)}{\partial X_{ij}^2} = \frac{1}{\tau}\cdot\frac{\mathrm{Tr}(\mathbb{1}\cdot P) - N_{ij}^2}{\mathrm{Tr}(\mathbb{1}\cdot P)^2}$$

and for the latter:

$$\frac{\partial^2\sigma(X,\tau)}{\partial X_{ij}\partial X_{l,q}} = -\frac{1}{\tau}\cdot\frac{-N_{ij}N_{lq}}{\mathrm{Tr}(\mathbb{1}\cdot P)^2}$$

Combining the two, we get the required result.

8.4 Proof of Lemma 4.4

Let us first prove convexity. By the definition of the Hessian, we want to prove

$$\mathrm{Tr}(\mathbb{1}\cdot P)\cdot y^\top\left(\mathtt{diag}(\mathtt{vec}(P)) - \frac{\mathtt{vec}(N)\mathtt{vec}(N)^\top}{\mathrm{Tr}(\mathbb{1}\cdot P)}\right)y \geq 0, \qquad \forall y\in\mathbb{R}^{mn}.$$

First, observe that $\mathrm{Tr}(\mathbb{1}\cdot P)\geq 0$ since each element of P is positive by definition. Second, for $P_{ij}\geq 0,\ \forall i,j$, it is obvious that $\frac{\mathtt{vec}(P)\mathtt{vec}(P)^\top}{\mathrm{Tr}(\mathbb{1}\cdot P)} \preceq \mathtt{diag}(\mathtt{vec}(P))$. Thus, what is left is to prove $y^\top\left(\mathtt{vec}(N)\mathtt{vec}(N)^\top\right)y \leq y^\top\left(\mathtt{vec}(P)\mathtt{vec}(P)^\top\right)y$, which is true since:

$$\begin{aligned} y^\top\left(\mathtt{vec}(N)\mathtt{vec}(N)^\top\right)y = \|y^\top\mathtt{vec}(N)\|_2^2 &= \sum_{i=1}^{mn}(y_i\cdot\mathtt{vec}(N)_i)^2 \leq \sum_{i=1}^{mn} y_i^2\cdot\mathtt{vec}(N)_i^2 \\ &\leq \sum_{i=1}^{mn} y_i^2\cdot\mathtt{vec}(P)_i^2 = \|y^\top\mathtt{vec}(P)\|_2^2 = y^\top\left(\mathtt{vec}(P)\mathtt{vec}(P)^\top\right)y, \end{aligned}$$

since $P_{ij}\geq N_{ij}$. Upper bounding the Hessian,

$$\begin{aligned} y^\top\nabla^2\sigma(X,\tau)y &= y^\top\left(\frac{1}{\tau}\cdot\frac{1}{\mathrm{Tr}(\mathbb{1}\cdot P)}\cdot\left(\mathtt{diag}(\mathtt{vec}(P)) - \frac{\mathtt{vec}(N)\mathtt{vec}(N)^\top}{\mathrm{Tr}(\mathbb{1}\cdot P)}\right)\right)y \\ &\leq y^\top\left(\frac{1}{\tau}\cdot\frac{1}{\mathrm{Tr}(\mathbb{1}\cdot P)}\cdot(\mathtt{diag}(\mathtt{vec}(P)))\right)y \\ &= \frac{\sum_{i=1}^{mn} y_i^2\cdot\mathtt{vec}(P)_i}{\tau\cdot\mathrm{Tr}(\mathbb{1}\cdot P)} \leq \frac{\sum_{i=1}^{mn}|y_i|^2\cdot(\sum_{i=1}^{mn}\mathtt{vec}(P)_i)}{\tau\cdot\mathrm{Tr}(\mathbb{1}\cdot P)} = \frac{\|y\|_2^2}{\tau}. \end{aligned}$$

This means that σ function is Lipschitz gradient continuous with constant $\frac{1}{\tau}$. To prove the set of inequalities of the lemma, we observe:

$$|X|_\infty \geq \sigma(X,\tau) \geq \tau\cdot\log\left(\frac{e^{|X|_\infty/\tau}}{2mn}\right) = |X|_\infty - \tau\log(2mn).$$

8.5 Proof of Theorem 5.1

Using Lemma 4.2, we bound $|M - U_T V_T^\top|_1$ as follows:

$$\begin{aligned} |M - U_T V_T^\top|_1 &\leq h(M - U_T V_T^\top,\tau) + mn\tau \\ &\leq h(M - U_T V_T^\top,\tau) + \frac{\lambda}{2}|U_T V_T^\top|_2^2 + mn\tau \end{aligned}$$

Define $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ such as $f(UV^\top) := h(M - UV^\top, \tau) + \frac{\lambda}{2}|UV^\top|_2^2$. Observe that f is λ-strongly convex with Lipscihtz continuous gradients with parameter $(\frac{2}{\tau} + \lambda)$. By Theorem 3.1, we know that:

$$f(U_T V_T^\top) - f(\widehat{U}^\star \widehat{V}^{\star\top}) \leq \frac{10 \cdot \textsc{Dist}(U_0, V_0; \widehat{X}_r^\star)^2}{\eta T}.$$

where $\textsc{Dist}(U_0, V_0; \widehat{X}_r^\star) \leq \frac{\sqrt{2}\cdot\sigma_r(\widehat{X}_r^\star)^{1/2}}{10\sqrt{\kappa}}$. Combining this bound with the above, we get:

$$|M - U_T V_T^\top|_1 \leq h(M - \widehat{U}^\star \widehat{V}^{\star\top}, \tau) + \frac{\lambda}{2}|\widehat{X}^\star|_2^2 + \frac{10 \cdot \textsc{Dist}(U_0, V_0; \widehat{X}_r^\star)^2}{\eta T} + mn\tau \tag{12}$$

We know from Lemma 4.2 that:

$$h(M - UV^\top, \tau) \leq |M - UV^\top|_1 \quad \Longrightarrow \quad h(M - UV^\top, \tau) + \frac{\lambda}{2}|UV^\top|_2^2 \leq |M - UV^\top|_1 + \frac{\lambda}{2}|UV^\top|_2^2$$

for every $U,\ V$. This further implies that:

$$\begin{aligned}
\min_{U,V}\left(h(M - UV^\top, \tau) + \frac{\lambda}{2}|UV^\top|_2^2\right) \leq \min_{U,V}\left(|M - UV^\top|_1 + \frac{\lambda}{2}|UV^\top|_2^2\right) \quad &\Rightarrow \\
h(M - \widehat{U}^\star \widehat{V}^{\star\top}, \tau) + \frac{\lambda}{2}|\widehat{U}^\star \widehat{V}^{\star\top}|_2^2 \overset{(i)}{\leq} \min_{U,V}\left(|M - UV^\top|_1 + \frac{\lambda}{2}|UV^\top|_2^2\right) & \\
\overset{(ii)}{\leq} |M - U^\star V^{\star\top}|_1 + \frac{\lambda}{2}|U^\star V^{\star\top}|_2^2 & \\
\overset{(iii)}{=} \mathrm{OPT} + \frac{\lambda}{2}|U^\star V^{\star\top}|_2^2 &
\end{aligned}$$

where (i) is due to the optimality of $\widehat{U}^\star, \widehat{V}^\star$ as the minimizer of $f(UV^\top) := h(M - UV^\top, \tau) + \frac{\lambda}{2}|UV^\top|_2^2$, (ii) is due to $U^\star, V^\star$ not being necessarily the minimizers of $\min_{U,V}\left(|M - UV^\top|_1 + \frac{\lambda}{2}|UV^\top|_2^2\right)$, and (iii) $\mathrm{OPT} := \min_{U,V}|M - UV^\top|_1 = |M - U^\star V^{\star\top}|_1$. Thus, (12) becomes:

$$|M - U_T V_T^\top|_1 \leq \mathrm{OPT} + \frac{\lambda}{2}|X^\star|_2^2 + \frac{10 \cdot \textsc{Dist}(U_0, V_0; X_r^\star)^2}{\eta T} + mn\tau$$

For $\varepsilon > 0$, setting $\tau = \frac{\varepsilon\cdot\mathrm{OPT}}{3mn}$ we observe that $mn\tau = \frac{\varepsilon\cdot\mathrm{OPT}}{3}$. Executing Algorithm 1 for $T \geq \frac{10\cdot\sigma_r(\widehat{X}_r^\star)}{50} \cdot \frac{3}{\eta\varepsilon\mathrm{OPT}}$, we can guarantee that $\frac{10\cdot\textsc{Dist}(U_0,V_0;\widehat{X}_r^\star)^2}{\eta T} \leq \frac{10\sigma_r(\widehat{X}^\star)}{50\eta\cdot\frac{3\cdot 10\cdot\sigma_r(\widehat{X}^\star)}{50\eta\varepsilon\mathrm{OPT}}} = \frac{\varepsilon\cdot\mathrm{OPT}}{3}$. Finally, setting $\lambda = \frac{2\varepsilon\cdot\mathrm{OPT}}{3|X^\star|_2^2}$, we obtain: $\frac{2\varepsilon\cdot\mathrm{OPT}}{6|X^\star|_2^2} \cdot |X^\star|_2^2 = \frac{\varepsilon\cdot\mathrm{OPT}}{3}$. Substituting the above in the main recursion, we get:

$$\begin{aligned}
|M - U_T V_T^\top|_1 &\leq \mathrm{OPT} + \frac{\lambda}{2}|X^\star|_2^2 + \frac{10 \cdot \textsc{Dist}(U_0, V_0; X_r^\star)^2}{\eta T} + mn\tau \\
&\leq \mathrm{OPT} + \frac{\varepsilon \cdot \mathrm{OPT}}{3} + \frac{\varepsilon \cdot \mathrm{OPT}}{3} + \frac{\varepsilon \cdot \mathrm{OPT}}{3} \\
&= (1 + \varepsilon) \cdot \mathrm{OPT}.
\end{aligned}$$

The number of iterations T required can be further analyzed to:

$$\begin{aligned}
T \geq \frac{10 \cdot \sigma_r(\widehat{X}_r^\star)}{50} \cdot \frac{3}{\eta\varepsilon\mathrm{OPT}} &\overset{(i)}{=} \frac{10 \cdot \sigma_r(\widehat{X}_r^\star)}{50} \cdot \frac{3 \cdot O(L)}{\varepsilon\mathrm{OPT}} \\
&\overset{(ii)}{=} \frac{10 \cdot \sigma_r(\widehat{X}_r^\star)}{50} \cdot \frac{3 \cdot O\left(\frac{1}{\tau} + \lambda\right)}{\varepsilon\mathrm{OPT}} \\
&\overset{(iii)}{=} \frac{10 \cdot \sigma_r(\widehat{X}_r^\star)}{50} \cdot \frac{3 \cdot O\left(\frac{3mn}{\varepsilon\mathrm{OPT}} + \frac{2\varepsilon\mathrm{OPT}}{3\|X^\star\|_2^2}\right)}{\varepsilon\mathrm{OPT}} \\
&= \frac{10 \cdot \sigma_r(\widehat{X}_r^\star)}{50} \cdot O\left(\frac{9mn}{(\varepsilon\mathrm{OPT})^2} + \frac{2}{\|X^\star\|_2^2}\right) \\
&= O\left(\sigma_r(\widehat{X}_r^\star) \cdot \left(\frac{mn}{(\varepsilon\mathrm{OPT})^2} + \frac{1}{\|X^\star\|_2^2}\right)\right)
\end{aligned}$$

where (i) is due to the definition of the step size that $\eta = O\left(\frac{1}{L}\right)$, (ii) is due to the definition $L = \frac{1}{\tau} + \lambda$, (iii) is obtained by substituting λ and τ.

8.6 Proof of Corollary 5.2

The proof is similar to that of Theorem 5.1. Using Lemma 4.4, we bound $|M - U_T V_T^\top|_\infty$ as follows:

$$\begin{aligned} |M - U_T V_T^\top|_\infty &\leq \sigma(U_T V_T^\top, \tau) + \tau \log(2mn) \\ &\leq \sigma(U_T V_T^\top, \tau) + \frac{\lambda}{2}|U_T V_T^\top|_2^2 + \tau \log(2mn) \end{aligned}$$

Following similar motions with Theorem 5.1, and setting $\tau = \frac{\varepsilon \cdot \mathrm{OPT}}{3\log(2mn)}$, and T and λ similar to the $p = 1$ case, we get:

$$|M - U_T V_T^\top|_\infty \leq (1 + \varepsilon) \cdot \mathrm{OPT}.$$

The number of iterations T required follow the same motions as the proof of Theorem 5.1, with a slight difference in the definition of τ.

9 CONNECTIONS WITH RELATED WORK

[10] considers probabilistic extensions of the PCA problem: starting with various generative probabilistic models, one obtains different matrix factorization objectives. The authors rely on the fundamental work of Csiszar and Tusnady [11], and propose an alternating minimization procedure; see also [49, 48].

[21, 45] show that the differences between many algorithms for matrix factorization can be viewed in terms of a small number of modeling choices. Their view unifies methods for Bregman co-clustering, LSI, non-negative matrix factorization, relational learning, to name a few.

While the bilinear factorization $UV^\top$ is common across different problems, there are cases where even a trilinear representation is more preferable, from an interpretation perspective. Having constraints over the factors is a another differentiation: An illustrative example of this case is that of matrix co-clustering where we are interested in $M \approx C_1 C_2^\top$, with C_1 and C_2 being matrices that denote the participation/indicator matrices. Our work is quite different to this type of factorizations (*i.e.*, with additional constraints on the factors); we defer the reader to [35, 18, 2, 53] for some recent developments on similar subjects.

Finally, there is a recent line of work on robust PCA that further focuses on identifying the (sparse) grossly corrupted elements in M; see [56, 6, 59, 31, 32, 8, 24, 57]. That line of work differs from our problem in that, our approach "models" the corruption through the penalization of the residual $M - UV^\top$ with an ℓ_1-norm, while in the aforementioned line of works, one optimizes over the residual $S = M - UV^\top$ in order to minimize the number of "active" corruptions. In that sense our model is "simpler" as we are only interested in identifying the low rank component.

10 SUPPORTIVE EXPERIMENTAL RESULTS

SVD		
	Time (sec.)	Error
Rank r	[min, mean, median]	
1	[2.63e-03, 1.10e-02, 1.08e-02]	[8.36e-01, 9.02e-01, 9.19e-01]
2	[3.44e-03, 5.58e-03, 4.25e-03]	[7.37e-01, 8.60e-01, 8.74e-01]
3	[4.08e-03, 8.55e-03, 6.67e-03]	[6.72e-01, 7.51e-01, 7.27e-01]
4	[2.59e-03, 7.73e-03, 4.47e-03]	[6.60e-01, 7.31e-01, 7.29e-01]
5	[2.59e-03, 3.69e-03, 3.63e-03]	[6.94e-01, 7.21e-01, 7.21e-01]
6	[2.52e-03, 3.40e-03, 3.11e-03]	[6.82e-01, 7.22e-01, 7.29e-01]
7	[2.44e-03, 3.21e-03, 3.29e-03]	[6.87e-01, 7.35e-01, 7.30e-01]
8	[2.43e-03, 3.58e-03, 3.32e-03]	[6.92e-01, 7.36e-01, 7.32e-01]
9	[2.50e-03, 3.01e-03, 2.97e-03]	[7.00e-01, 7.27e-01, 7.19e-01]
10	[1.96e-03, 2.70e-03, 2.84e-03]	[6.97e-01, 7.61e-01, 7.51e-01]

[17]		
	Time (sec.)	Error
Rank r	[min, mean, median]	
1	[6.81e-02, 2.24e-01, 2.28e-01]	[4.91e-01, 4.93e-01, 4.93e-01]
2	[1.55e-02, 2.75e-02, 2.31e-02]	[5.33e-01, 6.00e-01, 5.96e-01]
3	[2.42e-02, 5.89e-02, 4.59e-02]	[5.22e-01, 5.63e-01, 5.44e-01]
4	[2.69e-02, 4.61e-02, 4.04e-02]	[5.24e-01, 5.66e-01, 5.42e-01]
5	[4.67e-02, 3.36e-01, 1.48e-01]	[5.04e-01, 5.36e-01, 5.26e-01]
6	[6.72e-02, 6.24e-01, 1.34e-01]	[4.98e-01, 5.20e-01, 5.22e-01]
7	[5.46e-02, 8.91e-01, 5.47e-01]	[4.90e-01, 5.14e-01, 5.11e-01]
8	[1.36e-01, 1.66e+00, 5.39e-01]	[4.81e-01, 5.15e-01, 5.02e-01]
9	[1.90e-01, 2.91e+00, 2.56e+00]	[4.73e-01, 4.98e-01, 4.89e-01]
10	[2.30e-01, 9.60e+00, 4.25e+00]	[4.59e-01, 4.97e-01, 4.79e-01]

This work		
	Time (sec.)	Error
Rank r	[min, mean, median]	
1	[2.57e-02, 4.32e+01, 5.44e+01]	[4.99e-01, 5.82e-01, 5.01e-01]
2	[2.60e-02, 4.95e+01, 5.44e+01]	[5.04e-01, 5.49e-01, 5.07e-01]
3	[5.20e+01, 5.43e+01, 5.42e+01]	[5.06e-01, 5.10e-01, 5.10e-01]
4	[1.55e-02, 3.67e+01, 5.15e+01]	[5.05e-01, 5.90e-01, 5.10e-01]
5	[4.17e-02, 7.92e+01, 8.93e+01]	[5.07e-01, 5.33e-01, 5.13e-01]
6	[7.27e+01, 8.03e+01, 7.76e+01]	[5.02e-01, 5.08e-01, 5.09e-01]
7	[1.62e-02, 5.11e+01, 6.52e+01]	[5.08e-01, 5.84e-01, 5.08e-01]
8	[5.51e+01, 6.55e+01, 6.73e+01]	[4.95e-01, 5.09e-01, 5.02e-01]
9	[5.36e+01, 5.89e+01, 5.77e+01]	[4.78e-01, 5.06e-01, 5.06e-01]
10	[1.69e-02, 3.86e+01, 5.23e+01]	[4.69e-01, 5.94e-01, 4.75e-01]

Table 2: Attained objective function values and execution time. Table includes minimum, mean and median values for 10 Monte Carlo instances.