A Technical Lemmas

Lemma 4 (Theorem 1.1 in [40])**.** *There exists a constant K such that, for any n, m any $h \leq 2\log\max\{m,n\}$ and any $m \times n$ matrix $A = (a_{ij})$ where a_{ij} are i.i.d. symmetric random variables, the following inequality holds:*

$$\max\left\{\mathbb{E}\max_{1\leq i\leq m}\|a_{i\cdot}\|_2^h, \mathbb{E}\max_{1\leq j\leq n}\left\|a_{\cdot j}^h\right\|_2\right\} \leq \mathbb{E}\|A\|^h \leq K\left(\mathbb{E}\max_{1\leq i\leq m}\|a_{i\cdot}\|_2^h + \mathbb{E}\max_{1\leq j\leq n}\left\|a_{\cdot j}^h\right\|_2\right).$$

Lemma 5 (Symmetrization, Lemma 6.3 in [30])**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex. Then, for any finite sequence $\{t_i\}$ of independent mean zero random variables in B such that for every i $\mathbb{E}\left[F(\|t_i\|_2)\right] < \infty$, then*

$$\mathbb{E}\left[F\left(\frac{1}{2}\left\|\sum \xi_i t_i\right\|_2\right)\right] \leq \mathbb{E}\left[F\left(\left\|\sum t_i\right\|_2\right)\right] \leq \mathbb{E}\left[F\left(2\left\|\sum \xi_i t_i\right\|_2\right)\right],$$

where $\{\xi_i\}$ are i.i.d. Rademacher random variables.

Lemma 6 (Contraction, Theorem 4.12 in [30])**.** *Let $F : \mathbb{R}_+ \to \mathbb{R}_+$ be convex and increasing. Let $\psi_i : \mathbb{R} \to \mathbb{R}$ be contraction such that $\psi_i(0) = 0$. Then it holds that*

$$\mathbb{E}\left[F\left(\frac{1}{2}\sup_{t_1,\ldots t_N}\left|\sum_{i=1}^N \xi_i\psi_i(t_i)\right|\right)\right] \leq \mathbb{E}\left[F\left(\sup_{t_1,\ldots t_N}\left|\sum_{i=1}^N \xi_i t_i\right|\right)\right],$$

where $\{\xi_i\}$ are i.i.d. Rademacher random variables.

Lemma 7 (Lemma 2 in [17])**.** *Let f be a differentiable function and assume $\max\left\{\|M\|_\infty, \left\|\widehat{M}\right\|_\infty\right\} \leq \alpha$. Then*

$$d_H^2\left(f(M), f(\widehat{M})\right) \geq \inf_{|x|\leq\alpha}\frac{(f'(x))^2}{8f(x)(1-f(x))}\frac{\left\|M-\widehat{M}\right\|_F^2}{d_1 d_2}.$$

Lemma 8 (Lemma 4 in [17])**.** *Suppose that $x, y \in (0,1)$. Then*

$$D(x||y) \leq \frac{(x-y)^2}{y(1-y)}.$$

Lemma 9 (Lemma 3 in [17])**.** *Let $\mathcal{K}$ be the set of matrices that satisfy (A2) and (A3). Let $0 < \nu \leq 1$ be a scalar such that $r\nu^{-2}$ is an integer that is not larger than d_1. Then there exists a subset $\mathcal{X} \subset \mathcal{K}$ with the following properties:*

1. $|\mathcal{X}| \geq \exp\left(\frac{rd_2}{16\nu^2}\right)$.
2. $\forall\ X \in \mathcal{X}$, $|X_{ij}| = \alpha\nu$.
3. $\forall\ X, \widetilde{X} \in \mathcal{X}$ *with* $X \neq \widetilde{X}$, $\left\|X - \widetilde{X}\right\|_F^2 > \frac{1}{2}\alpha^2\nu^2 d_1 d_2$.

B Proof for Main Results

Recall the observation model: $M \in \mathbb{R}^{d_1\times d_2}$ is the true low-rank matrix and $\Omega \subset [d_1]\times[d_2]$ is the index set of entries we observed. $Y \in \mathbb{R}^{d_1\times d_2}$ is the binary matrix determined by M: for all $(i,j) \in \Omega$,

$$Y_{ij} = \begin{cases} +1, & \text{with probability } f(M_{ij}), \\ -1, & \text{with probability } 1 - f(M_{ij}). \end{cases}$$

In the setting of symmetric noise, the observation $Y'_{ij} = \delta_{ij}Y_{ij}$ where δ_{ij} are i.i.d. and

$$\delta_{ij} = \begin{cases} +1, & \text{with probability } 1-\tau, \\ -1, & \text{with probability } \tau, \end{cases}$$

where $\tau \in (0, 1/2)$ itself can be a random variable. Therefore, conditioning on τ, we observe

$$\Pr\left(Y'_{ij} = 1 \mid \tau\right) = (1-\tau) f(M_{ij}) + \tau(1 - f(M_{ij})).$$

Case 1. If τ is a discrete random variable, say

$$\Pr\left(\tau = \tau_k\right) = p_k,\ 1 \leq k \leq s,$$

then it is easy to see that

$$\begin{aligned}
\Pr\left(Y'_{ij} = 1\right) &= \sum_{k=1}^{s} \Pr\left(Y'_{ij} = 1,\ \tau = \tau_k\right) \\
&= \sum_{k=1}^{s} \Pr\left(Y'_{ij} = 1 \mid \tau = \tau_k\right) \cdot \Pr\left(\tau = \tau_k\right) \\
&= \sum_{k=1}^{s} p_k \Big[(1-\tau_k) f(M_{ij}) + \tau_k (1 - f(M_{ij}))\Big].
\end{aligned}$$

Denote

$$g(x) = \sum_{k=1}^{s} p_k \Big[(1-\tau_k) f(x) + \tau_k (1 - f(x))\Big] = (1 - 2\,\mathbb{E}[\tau]) f(x) + \mathbb{E}[\tau].$$

We have

$$Y'_{ij} = \begin{cases} +1, & \text{with probability } g(M_{ij}), \\ -1, & \text{with probability } 1 - g(M_{ij}). \end{cases}$$

Case 2. If τ is a continuous random variable with probability density function (pdf) $h_\tau(t)$, then we have

$$\begin{aligned}
\Pr\left(Y'_{ij} = 1\right) &= \int_t h_{Y,\tau}(Y'_{ij} = 1, t) dt \\
&= \int_t h_{Y|\tau}(Y'_{ij} = 1 \mid t) h_\tau(t) dt \\
&= \int_t h_\tau(t) \Big[(1-t) f(M_{ij}) + t(1 - f(M_{ij}))\Big] dt,
\end{aligned}$$

where $h_{Y,\tau}(y,t)$ is the joint pdf of Y_{ij} and τ, and $h_{Y|\tau}(y \mid t)$ is the conditional pdf. Thus, define

$$g(x) = \int_t h_\tau(t) \Big[(1-t) f(x) + t(1 - f(x))\Big] dt = (1 - 2\,\mathbb{E}[\tau]) f(x) + \mathbb{E}[\tau].$$

We again have

$$Y'_{ij} = \begin{cases} +1, & \text{with probability } g(M_{ij}), \\ -1, & \text{with probability } 1 - g(M_{ij}). \end{cases}$$

Hence, the maximum likelihood estimator is given as follows:

$$\widehat{M} = \arg\max_{X}\ L_{\Omega, Y'}(X), \quad \text{s.t. } \left\|X\right\|_* \leq \alpha\sqrt{r d_1 d_2},\ \left\|X\right\|_\infty \leq \gamma,$$

where

$$L_{\Omega, Y'}(X) := \sum_{(i,j)\in\Omega} \left(\mathbf{1}_{\{Y_{ij}=1\}} \log g(X_{ij}) + \mathbf{1}_{\{Y_{ij}=-1\}} \log(1 - g(X_{ij}))\right).$$

For the sake of a principled analysis, we will treat $g(x)$ as a general function at this point. Associated with the function $g(x)$ are two quantities:

$$\rho_\gamma^+ := \sup_{|x|\leq\gamma} \frac{|g'(x)|}{g(x)(1-g(x))}, \quad \rho_\gamma^- := \sup_{|x|\leq\gamma} \frac{g(x)(1-g(x))}{(g'(x))^2}.$$

We will use several kinds of distances in the proof. The first one is Hellinger distance that is given by

$$d_H^2(p,q) := (\sqrt{p}-\sqrt{q})^2 + (\sqrt{1-p}+\sqrt{1-q})^2, \quad \forall\ 0 \leq p,\ q \leq 1.$$

Extending it to the matrix, we write

$$d_H^2(P,Q) := \frac{1}{d_1 d_2} \sum_{i,j} d_H^2(P_{ij}, Q_{ij}),$$

where $P, Q \in \mathbb{R}^{d_1\times d_2}$ and the entries therein are between 0 and 1.

For two probability distributions $\mathcal{P}$ and $\mathcal{Q}$ on a finite set A, the Kullback-Leibler (KL) divergence is defined as

$$D(\mathcal{P}||\mathcal{Q}) = \sum_{x\in A} \mathcal{P}(x) \log \frac{\mathcal{P}(x)}{\mathcal{Q}(x)}.$$

With a slight abuse, we write for two scalars $p, q \in [0,1]$

$$D(p||q) = p\log\frac{p}{q} + (1-p)\log\frac{1-p}{1-q},$$

and for two matrices $P, Q \in [0,1]^{d_1\times d_2}$,

$$D(P||Q) = \frac{1}{d_1 d_2}\sum_{i,j} D(P_{ij}||Q_{ij}).$$

Throughout the proof, we will work with a shifted MLE, i.e.

$$\begin{aligned}\bar{L}_{\Omega,Y'}(X) := L_{\Omega,Y'}(X) - L_{\Omega,Y'}(0) &= \sum_{(i,j)\in\Omega} \left(\mathbf{1}_{\{Y_{ij}=1\}} \log\frac{g(X_{ij})}{g(0)} + \mathbf{1}_{\{Y_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(0)}\right)\\ &= \sum_{i,j} \mathbf{1}_{\{(i,j)\in\Omega\}} \left(\mathbf{1}_{\{Y_{ij}=1\}} \log\frac{g(X_{ij})}{g(0)} + \mathbf{1}_{\{Y_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(0)}\right). \end{aligned} \tag{12}$$

B.1 Proof for Lemma 3

Proof. Using the Markov's inequality, we have for any $\theta > 0$,

$$\begin{aligned}&\Pr\left(\sup_{X\in\mathcal{S}} \left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right| \geq \mathrm{C}_0 \alpha\rho_\gamma^+ \sqrt{r}\sqrt{n(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)\\ &\qquad\leq \frac{\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right|^\theta\right]}{\left(\mathrm{C}_0\alpha\rho_\gamma^+\sqrt{r}\sqrt{n(d_1+d_2)+d_1d_2\log(d_1d_2)}\right)^\theta}.\end{aligned} \tag{13}$$

We bound the numerator above. Recall that

$$\bar{L}_{\Omega,Y'}(X) = \sum_{i,j} \mathbf{1}_{\{(i,j)\in\Omega\}} \left(\mathbf{1}_{\{Y'_{ij}=1\}} \log\frac{g(X_{ij})}{g(0)} + \mathbf{1}_{\{Y'_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(0)}\right).$$

Let the random variable

$$\tilde{t}_{ij} = \mathbf{1}_{\{(i,j)\in\Omega\}} \left(\mathbf{1}_{\{Y'_{ij}=1\}} \log\frac{g(X_{ij})}{g(0)} + \mathbf{1}_{\{Y'_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(0)}\right),$$

and let

$$t_{ij} = \tilde{t}_{ij} - \mathbb{E}\,\tilde{t}_{ij}.$$

Then

$$\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X) = \sum_{i,j} t_{ij}.$$

Note that $\{t_{ij}\}$ are i.i.d. random variables with zero mean. The function $F(t) = \sup t^\theta$ is convex for $\theta \geq 1$, and $\mathbb{E}\,F(|t_{ij}|)$ is finite for all $(i,j) \in [d_1] \times [d_2]$. Hence, we can apply Lemma 5 to obtain

$$\begin{aligned}
&\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right|^\theta\right] \\
&\leq 2^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\sum_{i,j} \xi_{ij}\mathbf{1}_{\{(i,j)\in\Omega\}}\left(\mathbf{1}_{\{Y'_{ij}=1\}}\log\frac{g(X_{ij})}{g(0)} + \mathbf{1}_{\{Y'_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(0)}\right)\right|^\theta\right],
\end{aligned}$$

where $\{\xi_{ij}\}$ are i.i.d. Rademacher random variables. Now observe that due to the construction of ρ_γ^+, both $\frac{1}{\rho_\gamma^+}\log\frac{g(x)}{g(0)}$ and $\frac{1}{\rho_\gamma^+}\log\frac{1-g(x)}{1-g(0)}$ are contractions and vanish at $x=0$. Thereby, using Lemma 6 we have

$$\begin{aligned}
&\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right|^\theta\right] \\
&\leq (4\rho_\gamma^+)^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\sum_{i,j} \xi_{ij}\mathbf{1}_{\{(i,j)\in\Omega\}}\left(\mathbf{1}_{\{Y'_{ij}=1\}}X_{ij} - \mathbf{1}_{\{Y'_{ij}=-1\}}X_{ij}\right)\right|^\theta\right] \\
&= (4\rho_\gamma^+)^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\sum_{i,j} \xi_{ij}\mathbf{1}_{\{(i,j)\in\Omega\}}Y'_{ij}X_{ij}\right|^\theta\right].
\end{aligned}$$

With a simple algebra, we have

$$\Pr(\xi_{ij}Y'_{ij}=1) = \Pr(\xi_{ij}=1, Y'_{ij}=1) + \Pr(\xi_{ij}=-1, Y'_{ij}=-1) = \frac{1}{2}\left(\Pr(Y'_{ij}=1) + \Pr(Y'_{ij}=-1)\right) = \frac{1}{2},$$

which implies that the distribution of $\xi_{ij}Y'_{ij}$ is the same as that of ξ_{ij} for all $(i,j) \in [d_1]\times[d_2]$. Thus, by denoting Δ_Ω the matrix such that its (i,j)-th element is 1 if $(i,j)\in\Omega$ and 0 otherwise, and $\Xi = (\xi_{ij})$, it follows that

$$\begin{aligned}
\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right|^\theta\right] &\leq (4\rho_\gamma^+)^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\sum_{i,j}\xi_{ij}\mathbf{1}_{\{(i,j)\in\Omega\}}X_{ij}\right|^\theta\right] \\
&= (4\rho_\gamma^+)^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left|\langle \Delta_\Omega \circ \Xi, X\rangle\right|^\theta\right] \\
&\leq (4\rho_\gamma^+)^\theta\,\mathbb{E}\left[\sup_{X\in\mathcal{S}} \left\|\Delta_\Omega\circ\Xi\right\|^\theta \left\|X\right\|_*^\theta\right] \\
&\leq (\alpha\sqrt{rd_1d_2})^\theta\,\mathbb{E}\left[\left\|\Delta_\Omega\circ\Xi\right\|^\theta\right]. \qquad (14)
\end{aligned}$$

Above, the last inequality follows from the nuclear norm constraint we imposed in the MLE estimator. Note that the (i,j)-th entry of the matrix $\Delta_\Omega\circ\Xi$ is given by $\mathbf{1}_{\{(i,j)\in\Omega\}}\xi_{ij}$, which are i.i.d. symmetric random variables. Thus, Lemma 4 implies that

$$\begin{aligned}
\mathbb{E}\left[\left\|\Delta_\Omega\circ\Xi\right\|^\theta\right] &\leq \mathrm{C}\left(\mathbb{E}\max_{1\leq i\leq d_1}\left(\sum_{j=1}^{d_2}(\xi_{ij}\Delta_{ij})^2\right)^{\theta/2} + \mathbb{E}\max_{1\leq j\leq d_2}\left(\sum_{i=1}^{d_1}(\xi_{ij}\Delta_{ij})^2\right)^{\theta/2}\right) \\
&= \mathrm{C}\left(\mathbb{E}\max_{1\leq i\leq d_1}\left(\sum_{j=1}^{d_2}\Delta_{ij}\right)^{\theta/2} + \mathbb{E}\max_{1\leq j\leq d_2}\left(\sum_{i=1}^{d_1}\Delta_{ij}\right)^{\theta/2}\right). \qquad (15)
\end{aligned}$$

Fix i. By Bernstein's inequality, for all $t > 0$,

$$\Pr\left(\left|\sum_{j=1}^{d_2}\left(\Delta_{ij} - \frac{n}{d_1 d_2}\right)\right| > t\right) \leq 2\exp\left(\frac{-t^2/2}{n/d_1 + t/3}\right).$$

When $t \geq \frac{6n}{d_1}$, the above reduces to

$$\Pr\left(\left|\sum_{j=1}^{d_2}\left(\Delta_{ij} - \frac{n}{d_1 d_2}\right)\right| > t\right) \leq 2\exp(-t).$$

Suppose that $W_1, \ldots, W_{d_1}$ are i.i.d. exponential random variables with pdf $\exp(-t)$. Then it follows that

$$\Pr\left(\left|\sum_{j=1}^{d_2}\left(\Delta_{ij} - \frac{n}{d_1 d_2}\right)\right| > t\right) \leq 2\Pr(W_i \geq t).$$

On the other hand, we have

$$\begin{aligned}
\left(\mathbb{E}\max_{1\leq i\leq d_1}\left(\sum_{j=1}^{d_2}\Delta_{ij}\right)^{\theta/2}\right)^{1/\theta} &\leq \sqrt{\frac{n}{d_1}} + \left(\mathbb{E}\max_{1\leq i\leq d_1}\left|\sum_{j=1}^{d_2}(\Delta_{ij} - \frac{n}{d_1 d_2})\right|^{\theta/2}\right)^{1/\theta} \\
&\overset{\zeta_1}{=} \sqrt{\frac{n}{d_1}} + \left(\int_0^{+\infty}\Pr\left(\max_{1\leq i\leq d_1}\left|\sum_{j=1}^{d_2}(\Delta_{ij} - \frac{n}{d_1 d_2})\right|^{\theta/2} \geq t\right)dt\right)^{1/\theta} \\
&\leq \sqrt{\frac{n}{d_1}} + \left(\left(\frac{6n}{d_1}\right)^{\theta/2} + \int_{(6n/d_1)^{\theta/2}}^{+\infty}\Pr\left(\max_{1\leq i\leq d_1}\left|\sum_{j=1}^{d_2}(\Delta_{ij} - \frac{n}{d_1 d_2})\right|^{\theta/2} \geq t\right)dt\right)^{1/\theta} \\
&\leq \sqrt{\frac{n}{d_1}} + \left(\left(\frac{6n}{d_1}\right)^{\theta/2} + 2\int_{(6n/d_1)^{\theta/2}}^{+\infty}\Pr\left(\max_{1\leq i\leq d_1}W_i^{\theta/2} \geq t\right)dt\right)^{1/\theta} \\
&\overset{\zeta_2}{\leq} \sqrt{\frac{n}{d_1}} + \left(\left(\frac{6n}{d_1}\right)^{\theta/2} + 2\,\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2}\right)^{1/\theta} \\
&\leq (1+\sqrt{6})\sqrt{\frac{n}{d_1}} + 2^{1/\theta}\left(\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2}\right)^{1/\theta}.
\end{aligned}$$

Here, ζ_1 and ζ_2 use the identity $\mathbb{E}\,x = \int_0^{+\infty}\Pr(x \geq t)dt$ for any positive random variable x. It remains to bound $\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2}$. Using the fact that W_i is exponential, we have

$$\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2} \leq \left|\max_{1\leq i\leq d_1}W_i - \log d_1\right|^{\theta/2} + \log^{\theta/2} d_1 \leq 2((\theta/2)!) + \log^{\theta/2} d_1 \leq 2(\theta/2)^{\theta/2} + \log^{\theta/2} d_1,$$

where we apply Stirling's approximation in the last inequality. Thus,

$$2^{1/\theta}\left(\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2}\right)^{1/\theta} \leq 2^{1/\theta}\left(\sqrt{\log d_1} + 2^{1/\theta}\sqrt{\theta/2}\right).$$

Picking $\theta = 2\log(d_1 + d_2)$ gives

$$2^{1/\theta}\left(\mathbb{E}\max_{1\leq i\leq d_1}W_i^{\theta/2}\right)^{1/\theta} \leq (2+\sqrt{2})\sqrt{\log(d_1+d_2)}.$$

Putting pieces together, we obtain

$$\left(\mathbb{E}\max_{1\le i\le d_1}\left(\sum_{j=1}^{d_2}\Delta_{ij}\right)^{\theta/2}\right)^{1/\theta}\le(1+\sqrt{6})\sqrt{\frac{n}{d_1}}+(2+\sqrt{2})\sqrt{\log(d_1+d_2)}.$$

Likewise, we can show that

$$\left(\mathbb{E}\max_{1\le j\le d_2}\left(\sum_{i=1}^{d_1}\Delta_{ij}\right)^{\theta/2}\right)^{1/\theta}\le(1+\sqrt{6})\sqrt{\frac{n}{d_2}}+(2+\sqrt{2})\sqrt{\log(d_1+d_2)}.$$

Note that $\sqrt{x}$ is a concave function. Hence, Jensen's inequality implies that (15) can be bounded as follows:

$$\begin{aligned}\left(\mathbb{E}\left[\left\|\Delta_\Omega\circ\Xi\right\|^\theta\right]\right)^{1/\theta}&\le \mathrm{C}^{1/\theta}\left((1+\sqrt{6})\sqrt{\frac{2n(d_1+d_2)}{d_1d_2}}+(2+\sqrt{2})\sqrt{\log(d_1+d_2)}\right)\\&\le \mathrm{C}^{1/\theta}2(1+\sqrt{6})\sqrt{\frac{n(d_1+d_2)+d_1d_2\log(d_1+d_2)}{d_1d_2}}.\end{aligned}$$

Plugging this back to (14), we have

$$\left(\mathbb{E}\left[\sup_{X\in\mathcal{S}}\left|\bar{L}_{\Omega,Y'}(X)-\mathbb{E}\,\bar{L}_{\Omega,Y'}(X)\right|^\theta\right]\right)^{1/\theta}\le \mathrm{C}^{1/\theta}8(1+\sqrt{6})\alpha\rho_\gamma^+\sqrt{r}\sqrt{n(d_1+d_2)+d_1d_2\log(d_1+d_2)}.$$

Therefore, (13) is upper bounded by

$$\mathrm{C}\left(\frac{8(1+\sqrt{6})}{\mathrm{C}_0}\right)^{2\log(d_1+d_2)}\le\frac{\mathrm{C}}{d_1+d_2},$$

as soon as we choose $\mathrm{C}_0\ge 8(1+\sqrt{6})\sqrt{e}$. □

B.2 Proof for Theorem 1

We need the following result in our proof.

Proposition 10. *Assume same conditions as in Theorem 1 but with a slightly more general assumption that* $\|M\|_\infty\le\gamma$ *in place of* $\|M\|_\infty\le\alpha$. *Then, with probability at least* $1-\mathrm{C}_1/(d_1+d_2)$, *the follows holds:*

$$d_H^2(g(\widehat{M}),g(M))\le \mathrm{C}_2\rho_\gamma^+\alpha\sqrt{\frac{r(d_1+d_2)}{n}}\sqrt{1+\frac{(d_1+d_2)\log(d_1d_2)}{n}},$$

where C_1 *and* C_2 *are absolute constants.*

Proof. For any matrix $X\in\mathbb{R}^{d_1\times d_2}$, we have

$$\begin{aligned}\mathbb{E}\left[\bar{L}_{\Omega,Y'}(X)-\bar{L}_{\Omega,Y'}(M)\right]&=\mathbb{E}\left[L_{\Omega,Y'}(X)-L_{\Omega,Y'}(M)\right]\\&=\mathbb{E}\left[\sum_{i,j}\mathbf{1}_{\{(i,j)\in\Omega\}}\left(\mathbf{1}_{\{Y'_{ij}=1\}}\log\frac{g(X_{ij})}{g(M_{ij})}+\mathbf{1}_{\{Y'_{ij}=-1\}}\log\frac{1-g(X_{ij})}{1-g(M_{ij})}\right)\right]\\&=\mathbb{E}\left[\sum_{i,j}\frac{n}{d_1d_2}\left(g(M_{ij})\log\frac{g(X_{ij})}{g(M_{ij})}+(1-g(M_{ij}))\log\frac{1-g(X_{ij})}{1-g(M_{ij})}\right)\right]\\&=-nD(g(M)\|g(X)).\end{aligned}\tag{16}$$

On the other hand, for the optimum $\widehat{M}$, it holds that

$$\begin{aligned}\bar{L}_{\Omega,Y'}(\widehat{M}) - \bar{L}_{\Omega,Y'}(M) = &\ \mathbb{E}\left[\bar{L}_{\Omega,Y'}(\widehat{M}) - \bar{L}_{\Omega,Y'}(M)\right] + \left(\bar{L}_{\Omega,Y'}(\widehat{M}) - \mathbb{E}\left[\bar{L}_{\Omega,Y'}(\widehat{M})\right]\right) \\ &+ \left(\mathbb{E}\left[\bar{L}_{\Omega,Y'}(M) - \bar{L}_{\Omega,Y'}(M)\right)\right. \\ \leq &\ \mathbb{E}\left[\bar{L}_{\Omega,Y'}(X) - \bar{L}_{\Omega,Y'}(M)\right] + 2\sup_{X\in\mathcal{S}}\left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\left[\bar{L}_{\Omega,Y'}(X)\right]\right|,\end{aligned}$$

where we recall that $\mathcal{S}$ was defined in Lemma 3. Since $\widehat{M}$ also maximizes $\bar{L}_{\Omega,Y'}(X)$, we obtain

$$-\mathbb{E}\left[\bar{L}_{\Omega,Y'}(X) - \bar{L}_{\Omega,Y'}(M)\right] \leq 2\sup_{X\in\mathcal{S}}\left|\bar{L}_{\Omega,Y'}(X) - \mathbb{E}\left[\bar{L}_{\Omega,Y'}(X)\right]\right|.$$

This together with (16) and Lemma 3 imply that

$$D(g(M)||g(\widehat{M})) \leq 2\mathrm{C}_0\alpha_0\rho_\gamma^+\sqrt{\frac{r(d_1+d_2)}{n}}\sqrt{1+\frac{(d_1+d_2)\log(d_1d_2)}{n}}$$

holds with probability at least $1-\mathrm{C}_1/(d_1+d_2)$. Since the Hellinger distance is upper bounded by the KL divergence, we complete the proof. □

Now we are in the position to prove Theorem 1. In fact, Theorem 1 follows immediately from Prop. 10 and Lemma 7.

B.3 Proof for Theorem 2

Proof. Without loss of generality, suppose that $d_1 \leq d_2$. Let

$$\epsilon^2 = \min\left\{\frac{1}{1024}, \mathrm{C}\alpha\sqrt{\frac{\rho^-_{0.75\alpha}rd_2}{n}}\right\}.$$

Pick

$$\frac{4\sqrt{2}\epsilon}{\alpha} \leq \nu \leq \frac{8\epsilon}{\alpha}.$$

It is easy to see that

$$\frac{r\alpha^2}{64\epsilon^2} \leq \frac{r}{\nu^2} \leq \frac{r\alpha^2}{32\epsilon^2}.$$

The length of this interval is $\frac{r\alpha^2}{64\epsilon}$, which is larger than 1 since $\alpha \geq 1$, $r \geq 16$ and $\epsilon^2 \leq 1/1024$. Hence, it is possible to pick a proper ν such that $\frac{r}{\nu^2}$ is an integer. Also, the assumption that $\epsilon^2 \geq O(r\alpha^2/d_1)$ suggests $r/\nu^2 \leq d_1$. Hence we have found an appropriate ν for Lemma 9.

Let $\mathcal{X}'_{\alpha/2,\nu}$ be a set that satisfies all the properties in Lemma 9 with parameter $\alpha/2$. Let

$$\mathcal{X} = \left\{X' + \alpha\left(1-\frac{\nu}{2}\right)U : X' \in \mathcal{X}'_{\alpha/2,\nu}\right\},$$

where all the entries of U equal one.

First, we verify that each component in $\mathcal{X}$ satisfies $(A2)$ and $(A3)$. It is easy to see that for any $X \in \mathcal{X}$, $|X_{ij}|$ either equals α or $(1-\nu)\alpha$, i.e., $\|X\|_\infty \leq \alpha$ since $\nu < 1$. In addition,

$$\left\|X' + \alpha\left(1-\frac{\nu}{2}\right)U\right\|_* \leq \|X'\|_* + \alpha\left(1-\frac{\nu}{2}\right)\|U\|_* \leq \frac{\alpha}{2}\sqrt{rd_1d_2} + \alpha\left(1-\frac{\nu}{2}\right)\|U\|_F.$$

Since $\nu \in (0,1)$ and $r \geq 16$, we have $2-\nu \leq \sqrt{r}$, which together with $\|U\|_F = \sqrt{d_1d_2}$ imply that $\|X\|_* \leq \alpha\sqrt{rd_1d_2}$ for all $X \in \mathcal{X}$.

We prove the theorem by showing that its converse is false. That is, suppose that there exists an algorithm such that for any $M \in \mathcal{X}$ (which satisfies $(A2)$ and $(A3)$), with probability at least $1/4$, its output $\widehat{X}$ satisfies

$$\frac{1}{d_1 d_2} \left\| \widehat{X} - M \right\|_F^2 < \epsilon^2. \tag{17}$$

Let $X^* \in \mathcal{X}$ be the closest member to $\widehat{X}$. For any $\widetilde{X} \neq M \in \mathcal{X}$, it follows that

$$\left\| \widetilde{X} - \widehat{X} \right\|_F \geq \left\| \widetilde{X} - M \right\|_F - \left\| \widehat{X} - M \right\|_F > 2\epsilon\sqrt{d_1 d_2} - \epsilon\sqrt{d_1 d_2} = \epsilon\sqrt{d_1 d_2}, \tag{18}$$

where the last inequality follows from (17) and the fact that for any $X, \widetilde{X} \in \mathcal{X}$,

$$\left\| X - \widetilde{X} \right\|_F^2 \geq \frac{\alpha^2 \nu^2 d_1 d_2}{8} \geq 4 d_1 d_2 \epsilon^2.$$

The first inequality above uses the third property in Lemma 9 and the second inequality follows from our choice of ν.

On the other hand, since X^* is the closest one to $\widehat{X}$, we have

$$\left\| X^* - \widehat{X} \right\|_F \leq \left\| M - \widehat{X} \right\|_F \leq \epsilon\sqrt{d_1 d_2}. \tag{19}$$

Combining (18) and (19), we obtain

$$\left\| X^* - \widehat{X} \right\|_F < \left\| \widetilde{X} - \widehat{X} \right\|_F, \ \forall\ \widetilde{X} \neq M,$$

which implies $X^* = M$. Since (17) holds with probability at least $1/4$,

$$\Pr\left(X^* \neq M\right) \leq \frac{3}{4}. \tag{20}$$

From a variant of Fano's inequality,

$$\Pr(X^* \neq M) \geq 1 - \frac{1 + d_1 d_2 \max_{X \neq \widetilde{X}} D(Y'_\Omega | X \ || \ Y'_\Omega | \widetilde{X})}{\log |\mathcal{X}|} \tag{21}$$

Denote

$$D = d_1 d_2 D(Y'_\Omega | X \ || \ Y'_\Omega | \widetilde{X}) = \sum_{(i,j) \in \Omega} D(Y'_{ij} | X_{ij} \ || \ Y'_{ij} | \widetilde{X}_{ij}).$$

For each $(i,j) \in \Omega$, $D(Y'_{ij} | X_{ij} \ || \ Y'_{ij} | \widetilde{X}_{ij})$ is either 0, $D(g(\alpha) || g(\alpha'))$ or $D(g(\alpha) || g(\alpha'))$, where $\alpha' = (1-\nu)\alpha$ and we recall that X_{ij}, $\widetilde{X}_{ij}$ can only take value from $\{\alpha, \alpha'\}$. It thus follows from Lemma 8 that

$$D(Y'_{ij} | X_{ij} \ || \ Y'_{ij} | \widetilde{X}_{ij}) \leq \frac{(g(\alpha) - g(\alpha'))^2}{g(\alpha')(1 - g(\alpha'))},$$

since $\alpha' < \alpha$. Now using the mean value theorem, we obtain

$$D \leq n (g'(\theta))^2 \frac{(\alpha - \alpha')^2}{g(\alpha')(1 - g(\alpha'))}, \text{ for some } \theta \in [\alpha', \alpha].$$

As we assumed that $\nabla g(x)$ is decreasing in $(0, +\infty)$, we get

$$D \leq \frac{n (\nu\alpha)^2}{\rho^-_{\alpha'}} \leq \frac{64 n \epsilon^2}{\rho^-_{\alpha'}}.$$

Due to the construction, the cardinality of $\mathcal{X}$ equals to that of $\mathcal{X}'_{\alpha/2,\nu}$. Hence, combining (20) and (21), we can show

$$\frac{1}{4} \leq \frac{D+1}{\log|\mathcal{X}|} \leq \frac{16\nu^2}{rd_2}\left(\frac{64n\epsilon^2}{\rho^-_{\alpha'}}+1\right) \leq \frac{1024\epsilon^2}{\alpha^2 rd_2}\left(\frac{64n\epsilon^2}{\rho^-_{\alpha'}}+1\right). \tag{22}$$

Note that when $64n\epsilon^2 \leq \rho^-_{\alpha'}$, we have

$$\frac{1}{4} \leq 1024\frac{2048\epsilon^2}{\alpha^2 rd_2},$$

implying $\alpha^2 rd_2 \leq 8$ due to the definition of ϵ. This contradicts our assumption that $\alpha^2 rd_2 \geq \mathrm{C}_0$ if we specify $\mathrm{C}_0 > 8$.

When $64n\epsilon^2 > \rho^-_{\alpha'}$, then (22) suggests

$$\frac{1}{4} \leq \frac{1024 \times 128 \times n\epsilon^4}{\rho^-_{\alpha'}\alpha^2 rd_2},$$

which gives

$$\epsilon^2 > \frac{\alpha\sqrt{\rho^-_{\alpha'}}}{1024}\sqrt{\frac{rd_2}{n}}.$$

Picking $\mathrm{C}_2 = 1/1024$ in the definition of ϵ and noting $\rho^-_{\alpha'} \geq \rho^-_{0.75\alpha}$ yields a contradiction.

Therefore, (17) fails to hold with probability at least $3/4$.

$\square$