Matrix reconstruction with the local max norm

Rina Foygel
Department of Statistics
Stanford University
rinafb@stanford.edu

Nathan Srebro
Toyota Technological Institute at Chicago
nati@ttic.edu

Ruslan Salakhutdinov
Dept. of Statistics and Dept. of Computer Science University of Toronto
rsalakhu@utstat.toronto.edu

Supplementary Materials

A Proof of Theorem 1

Special case: element-wise upper bounds First, we assume that the general result is true, i.e.

$$2\left\|X\right\|_{(\mathcal{R},\mathcal{C})} = \inf_{AB^\top=X}\left(\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i\left\|A_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j\left\|B_{(j)}\right\|_2^2\right)\ , \tag{1}$$

and prove the result in the special case, where

$$\mathcal{R} = \{\mathbf{r}\in\Delta_{[n]} : \mathbf{r}_i \le R_i\ \forall i\} \text{ and } \mathcal{C} = \{\mathbf{c}\in\Delta_{[m]} : \mathbf{c}_j \le C_j\ \forall j\}\ .$$

Using strong duality for linear programs, we have

$$\begin{aligned}\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i\left\|A_{(i)}\right\|_2^2 &= \sup_{\mathbf{r}\in\mathbb{R}^n_+}\left\{\sum_i \mathbf{r}_i\left\|A_{(i)}\right\|_2^2\ :\ \mathbf{r}_i\le R_i,\ \sum_i \mathbf{r}_i = 1\right\}\\ &= \inf_{a\in\mathbb{R},a_1\in\mathbb{R}^n_+}\left\{a + R^\top a_1\ :\ a + a_{1i} \ge \left\|A_{(i)}\right\|_2^2\ \forall i\right\}\ .\end{aligned}$$

In this last line, if we fix a and want to minimize over $a_1 \in \mathbb{R}^n_+$, it is clear that the infimum is obtained by setting $a_{1i} = (\left\|A_{(i)}\right\|_2^2 - a)_+$ for each i. This proves that

$$\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i\left\|A_{(i)}\right\|_2^2 = \inf_{a\in\mathbb{R}}\left\{a + \sum_i R_i\left(\left\|A_{(i)}\right\|_2^2 - a\right)_+\right\}\ .$$

Applying the same reasoning to the columns and plugging everything in to (1), we get

$$2\left\|X\right\|_{(\mathcal{R},\mathcal{C})} = \inf_{AB^\top=X,\ a,b\in\mathbb{R}}\left\{a + \sum_i R_i\left(\left\|A_{(i)}\right\|_2^2 - a\right)_+ + b + \sum_j C_j\left(\left\|B_{(j)}\right\|_2^2 - b\right)_+\right\}\ .$$

General factorization result In the proof sketch given in the main paper, we showed that

$$2\left\|X\right\|_{(\mathcal{R},\mathcal{C})} \le \inf_{AB^\top=X}\left(\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}A\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}}\left\|\mathbf{c}^{1/2}B\right\|_{\mathrm{F}}^2\right)\ .$$

We now want to prove the reverse inequality. Since $\left\|X\right\|_{(\mathcal{R},\mathcal{C})} = \left\|X\right\|_{(\overline{\mathcal{R}},\overline{\mathcal{C}})}$ by definition (where $\overline{\mathcal{S}}$ denotes the closure of a set $\mathcal{S}$), we can assume without loss of generality that $\mathcal{R}$ and $\mathcal{C}$ are both closed (and compact) sets.

First, we restrict our attention to a special case (the "positive case"), where we assume that for all $\mathbf{r} \in \mathcal{R}$ and all $\mathbf{c} \in \mathcal{C}$, $\mathbf{r}_i > 0$ and $\mathbf{c}_j > 0$ for all i and j. (We will treat the general case below.) Therefore, since $\|X\|_{\text{tr}(\mathbf{r},\mathbf{c})}$ is continuous as a function of $(\mathbf{r}, \mathbf{c})$ for any fixed X and since $\mathcal{R}$ and $\mathcal{C}$ are closed, we must have some $\mathbf{r}^\star \in \mathcal{R}$ and $\mathbf{c}^\star \in \mathcal{C}$ such that $\|X\|_{(\mathcal{R},\mathcal{C})} = \|X\|_{\text{tr}(\mathbf{r}^\star,\mathbf{c}^\star)}$, with $\mathbf{r}^\star_i > 0$ for all i and $\mathbf{c}^\star_j > 0$ for all j.

Next, let $UDV^\top = {\mathbf{r}^\star}^{1/2} \cdot X \cdot {\mathbf{c}^\star}^{1/2}$ be a singular value decomposition, and let $A^\star = {\mathbf{r}^\star}^{-1/2} U D^{1/2}$ and $B^\star = {\mathbf{c}^\star}^{-1/2} V D^{1/2}$. Then $A^\star {B^\star}^\top = X$, and

$$\left\|{\mathbf{r}^\star}^{1/2} A^\star\right\|_{\text{F}}^2 = \left\|UD^{1/2}\right\|_{\text{F}}^2 = \text{trace}(UDU^\top) = \text{trace}(D) = \|X\|_{\text{tr}(\mathbf{r}^\star,\mathbf{c}^\star)} = \|X\|_{(\mathcal{R},\mathcal{C})} \ .$$

Below, we will show that

$$\mathbf{r}^\star = \arg\max_{\mathbf{r}\in\mathcal{R}} \left\|\mathbf{r}^{1/2} A^\star\right\|_{\text{F}}^2 \ . \tag{2}$$

This will imply that $\|X\|_{(\mathcal{R},\mathcal{C})} = \sup_{\mathbf{r}\in\mathcal{R}} \left\|\mathbf{r}^{1/2} A^\star\right\|_{\text{F}}^2$, and following the same reasoning for $B^\star$, we will have proved

$$2\|X\|_{(\mathcal{R},\mathcal{C})} = \left(\sup_{\mathbf{r}\in\mathcal{R}} \left\|\mathbf{r}^{1/2} A^\star\right\|_{\text{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}} \left\|\mathbf{c}^{1/2} B^\star\right\|_{\text{F}}^2\right) \geq \inf_{AB^\top = X} \left(\sup_{\mathbf{r}\in\mathcal{R}} \left\|\mathbf{r}^{1/2} A\right\|_{\text{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}} \left\|\mathbf{c}^{1/2} B\right\|_{\text{F}}^2\right) \ ,$$

which is sufficient. It remains only to prove (2). Take any $\mathbf{r} \in \mathcal{R}$ with $\mathbf{r} \neq \mathbf{r}^\star$ and let $\mathbf{w} = \mathbf{r} - \mathbf{r}^\star$. We have

$$\left\|\mathbf{r}^{1/2} A\right\|_{\text{F}}^2 - \left\|{\mathbf{r}^\star}^{1/2} A\right\|_{\text{F}}^2 = \sum_i \mathbf{w}_i \left\|A_{(i)}\right\|_2^2 = \sum_i \frac{\mathbf{w}_i}{\mathbf{r}^\star_i} \cdot (UDU^\top)_{ii} \ ,$$

and it will be sufficient to prove that this quantity is ≤ 0. To do this, we first define, for any $t \in [0, 1]$,

$$f(t) := \sum_i \sqrt{1 + t \cdot \frac{\mathbf{w}_i}{\mathbf{r}^\star_i}} \cdot (UDU^\top)_{ii} = \text{trace}\left(\left(\frac{\mathbf{r}^\star + t\mathbf{w}}{\mathbf{r}^\star}\right)^{1/2} UDU^\top\right) \ .$$

Using the fact that $\text{trace}(\cdot) \leq \|\cdot\|_{\text{tr}}$ for all matrices, we have

$$\begin{aligned} f(t) &\leq \left\|\left(\frac{\mathbf{r}^\star + t\mathbf{w}}{\mathbf{r}^\star}\right)^{1/2} UDU^\top\right\|_{\text{tr}} = \left\|(\mathbf{r}^\star + t\mathbf{w})^{1/2} X {\mathbf{c}^\star}^{1/2} \cdot VU^\top\right\|_{\text{tr}} \\ &= \left\|(\mathbf{r}^\star + t\mathbf{w})^{1/2} X {\mathbf{c}^\star}^{1/2}\right\|_{\text{tr}} = \|X\|_{\text{tr}(\mathbf{r}^\star + t\mathbf{w}, \mathbf{c}^\star)} \leq \|X\|_{(\mathcal{R},\mathcal{C})} = \sum_i (UDU^\top)_{ii} = f(0) \ , \end{aligned}$$

where the last inequality comes from the fact that $\mathbf{r}^\star + t\mathbf{w} \in \mathcal{R}$ by convexity of $\mathcal{R}$. Therefore,

$$0 \geq \frac{d}{dt} f(t)\Big|_{t=0} = \frac{d}{dt}\left(\sum_i \sqrt{1 + t \cdot \frac{\mathbf{w}_i}{\mathbf{r}^\star_i}} \cdot (UDU^\top)_{ii}\right)\Big|_{t=0} = \frac{1}{2} \cdot \sum_i \frac{\mathbf{w}_i}{\mathbf{r}^\star_i} \cdot (UDU^\top)_{ii} \ ,$$

as desired. (Here we take the right-sided derivative, i.e. taking a limit as t approaches zero from the right, since $f(t)$ is only defined for $t \in [0, 1]$.) This concludes the proof for the positive case.

Next, we prove that the general factorization (1) hold in the general case, where we might have $\mathcal{R} \not\subset \mathbb{R}^n_{++}$ and/or $\mathcal{C} \not\subset \mathbb{R}^m_{++}$. If for any $i \in [n]$ we have $\mathbf{r}_i = 0$ for all $\mathbf{r} \in \mathcal{R}$, we can discard this row of X, and same for any $j \in [m]$. Therefore, without loss of generality, for all $i \in [n]$ there is some $\mathbf{r}^{(i)} \in \mathcal{R}$ with $\mathbf{r}^{(i)}_i > 0$. Taking a convex combination, $\mathbf{r}^+ = \frac{1}{n}\sum_i \mathbf{r}^{(i)} \in \mathcal{R}$, we have $\mathbf{r}^+ \in \mathcal{R} \cap \mathbb{R}^n_{++}$. Similarly, we can construct $\mathbf{c}^+ \in \mathcal{C} \cap \mathbb{R}^m_{++}$.

Fix any $\epsilon > 0$, and let $\delta = \min\{\min_i \mathbf{r}^+_i, \min_j \mathbf{c}^+_j\} \cdot \frac{\epsilon}{2(1+\epsilon)} > 0$, and define closed subsets

$$\mathcal{R}_0 = \left\{\mathbf{r} \in \mathcal{R} : \min_i \mathbf{r}_i \geq \delta\right\} \subseteq \mathcal{R} \text{ and } \mathcal{C}_0 = \left\{\mathbf{c} \in \mathcal{C} : \min_i \mathbf{c}_i \geq \delta\right\} \subseteq \mathcal{C} \ .$$

Since we know that the factorization result holds for the "positive case", we have

$$\begin{aligned} \inf_{AB^\top = X} &\left(\sup_{\mathbf{r}\in\mathcal{R}_0} \left\|\mathbf{r}^{1/2} A\right\|_{\text{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}_0} \left\|\mathbf{c}^{1/2} B\right\|_{\text{F}}^2\right) = 2\|X\|_{(\mathcal{R}_0,\mathcal{C}_0)} \\ &= 2 \sup_{\mathbf{r}\in\mathcal{R}_0, \mathbf{c}\in\mathcal{C}_0} \left\|\mathbf{r}^{1/2} X \mathbf{c}^{1/2}\right\|_{\text{tr}} \leq 2 \sup_{\mathbf{r}\in\mathcal{R}, \mathbf{c}\in\mathcal{C}} \left\|\mathbf{r}^{1/2} X \mathbf{c}^{1/2}\right\|_{\text{tr}} = 2\|X\|_{(\mathcal{R},\mathcal{C})} \ . \end{aligned}$$

Now choose any factorization $\tilde{A}\tilde{B}^\top = X$ such that

$$\left(\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}_0}\left\|\mathbf{c}^{1/2}\tilde{B}\right\|_{\mathrm{F}}^2\right) \le 2\sup_{\mathbf{r}\in\mathcal{R},\mathbf{c}\in\mathcal{C}}\left\|\mathbf{r}^{1/2}X\mathbf{c}^{1/2}\right\|_{\mathrm{tr}}(1+\epsilon/2)\ . \tag{3}$$

Next, we need to show that $\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}^2$ is not much larger than $\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}^2$ (and same for $\tilde{B}$). Choose any $\mathbf{r}'\in\mathcal{R}$, and let $\mathbf{r}'' = \left(1-\frac{\delta}{\min_i \mathbf{r}_i^+}\right)\mathbf{r}' + \left(\frac{\delta}{\min_i \mathbf{r}_i^+}\right)\mathbf{r}^+ \in \mathcal{R}$. Then

$$\min_i \mathbf{r}''_i \ge \left(\frac{\delta}{\min_i \mathbf{r}_i^+}\right)\min_i \mathbf{r}_i^+ = \delta\ ,$$

and so $\mathbf{r}''\in\mathcal{R}_0$. We also have $\mathbf{r}'_i \le \left(1-\frac{\delta}{\min_i \mathbf{r}_i^+}\right)^{-1}\mathbf{r}''_i$ for all i. Therefore,

$$\left\|\mathbf{r}'^{1/2}\tilde{A}\right\|_{\mathrm{F}} \le \left(1-\frac{\delta}{\min_i \mathbf{r}_i^+}\right)^{-1/2}\left\|\mathbf{r}''^{1/2}\tilde{A}\right\|_{\mathrm{F}} \le \left(1-\frac{\delta}{\min_i \mathbf{r}_i^+}\right)^{-1/2}\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}\ .$$

Since this is true for any $\mathbf{r}'\in\mathcal{R}$, applying the definition of δ, we have

$$\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}} \le \left(1-\frac{\delta}{\min_i \mathbf{r}_i^+}\right)^{-1/2}\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}} \le \left(\frac{1+\epsilon/2}{1+\epsilon}\right)^{-1/2}\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}\ .$$

Applying the same reasoning for $\tilde{B}$ and then plugging in the bound (3), we have

$$\begin{aligned}\inf_{AB^\top=X}\left(\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}A\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}}\left\|\mathbf{c}^{1/2}B\right\|_{\mathrm{F}}^2\right) &\le \left(\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}} + \sup_{\mathbf{c}\in\mathcal{C}}\left\|\mathbf{c}^{1/2}\tilde{B}\right\|_{\mathrm{F}}^2\right)\\ &\le \left(\frac{1+\epsilon/2}{1+\epsilon}\right)^{-1}\cdot\left(\sup_{\mathbf{r}\in\mathcal{R}_0}\left\|\mathbf{r}^{1/2}\tilde{A}\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}_0}\left\|\mathbf{c}^{1/2}\tilde{B}\right\|_{\mathrm{F}}^2\right)\\ &\le \left(\frac{1+\epsilon/2}{1+\epsilon}\right)^{-1}(1+\epsilon/2)\cdot 2\left\|X\right\|_{(\mathcal{R},\mathcal{C})} = (1+\epsilon)\cdot 2\left\|X\right\|_{(\mathcal{R},\mathcal{C})}\ .\end{aligned}$$

Since this analysis holds for arbitrary $\epsilon > 0$, this proves the desired result, that

$$\inf_{AB^\top=X}\left(\sup_{\mathbf{r}\in\mathcal{R}}\left\|\mathbf{r}^{1/2}A\right\|_{\mathrm{F}}^2 + \sup_{\mathbf{c}\in\mathcal{C}}\left\|\mathbf{c}^{1/2}B\right\|_{\mathrm{F}}^2\right) \le 2\left\|X\right\|_{(\mathcal{R},\mathcal{C})}\ .$$

B Proof of Theorem 2

We follow similar techniques as used by Srebro and Shraibman [1] in their proof of the analogous result for the max norm. We need to show that

$$\begin{aligned}&\mathrm{Conv}\left\{uv^\top : u\in\mathbb{R}^n, v\in\mathbb{R}^m, \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\} \subseteq \left\{X : \|X\|_{(\mathcal{R},\mathcal{C})}\le 1\right\} \subseteq\\ &\qquad K_G\cdot\mathrm{Conv}\left\{uv^\top : u\in\mathbb{R}^n, v\in\mathbb{R}^m, \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\}\ .\end{aligned}$$

For the left-hand inclusion, since $\|\cdot\|_{(\mathcal{R},\mathcal{C})}$ is a norm and therefore the constraint $\|X\|_{(\mathcal{R},\mathcal{C})}\le 1$ is convex, it is sufficient to show that $\left\|uv^\top\right\|_{(\mathcal{R},\mathcal{C})}\le 1$ for any $u\in\mathbb{R}^n, v\in\mathbb{R}^m$ with $\|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1$. This is a trivial consequence of the factorization result in Theorem 1.

Now we prove the right-hand inclusion. Grothendieck's Inequality states that, for any $Y\in\mathbb{R}^{n\times m}$ and for any dimension k,

$$\begin{aligned}&\sup\left\{\langle Y, UV^\top\rangle : U\in\mathbb{R}^{n\times k}, V\in\mathbb{R}^{m\times k}, \left\|U_{(i)}\right\|_2\le 1\ \forall i,\ \left\|V_{(j)}\right\|_2\le 1\ \forall j\right\}\\ &\qquad\le K_G\cdot\sup\left\{\langle Y, uv^\top\rangle : u\in\mathbb{R}^n, v\in\mathbb{R}^m, |u_i|\le 1\ \forall i,\ |v_j|\le 1\ \forall j\right\}\ ,\end{aligned}$$

where $K_G \in (1.67, 1.79)$ is Grothendieck's constant. We now extend this to a slightly more general form. Take any $a \in \mathbb{R}^n_+$ and $b \in \mathbb{R}^m_+$. Then, setting $\tilde{U} = \mathrm{diag}(a)^+ U$ and $\tilde{V} = \mathrm{diag}(b)^+ V$ (where M^+ is the pseudoinverse of M), and same for $\tilde{u}$ and $\tilde{v}$, we see that

$$\begin{aligned}
&\sup\left\{\langle Y, UV^\top\rangle : U \in \mathbb{R}^{n\times k}, V \in \mathbb{R}^{m\times k}, \left\|U_{(i)}\right\|_2 \le a_i \;\forall i,\; \left\|V_{(j)}\right\|_2 \le b_j \;\forall j\right\} \\
&= \sup\left\{\langle \mathrm{diag}(a)\cdot Y\cdot \mathrm{diag}(b), \tilde{U}\tilde{V}^\top\rangle : \tilde{U} \in \mathbb{R}^{n\times k}, \tilde{V} \in \mathbb{R}^{m\times k}, \left\|\tilde{U}_{(i)}\right\|_2 \le 1 \;\forall i,\; \left\|\tilde{V}_{(j)}\right\|_2 \le 1 \;\forall j\right\} \\
&\le K_G \cdot \sup\left\{\langle \mathrm{diag}(a)\cdot Y\cdot \mathrm{diag}(b), \tilde{u}\tilde{v}^\top\rangle : \tilde{u} \in \mathbb{R}^{n}, \tilde{v} \in \mathbb{R}^{m}, |\tilde{u}_i| \le 1 \;\forall i,\; |\tilde{v}_j| \le 1 \;\forall j\right\} \\
&= K_G \cdot \sup\left\{\langle Y, uv^\top\rangle : u \in \mathbb{R}^{n}, v \in \mathbb{R}^{m}, |u_i| \le a_i \;\forall i,\; |v_j| \le b_j \;\forall j\right\} . \quad (4)
\end{aligned}$$

Now take any $Y \in \mathbb{R}^{n\times m}$. Let $\|\cdot\|^*_{(\mathcal{R},\mathcal{C})}$ be the dual norm to the $(\mathcal{R},\mathcal{C})$-norm. To bound this dual norm of Y, we apply the factorization result of Theorem 1:

$$\begin{aligned}
\|Y\|^*_{(\mathcal{R},\mathcal{C})} &= \sup_{\|X\|_{(\mathcal{R},\mathcal{C})}\le 1} \langle Y, X\rangle \\
&= \sup_{U,V}\left\{\langle Y, UV^\top\rangle : \frac{1}{2}\left(\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2\right) \le 1\right\} \\
&\overset{(*)}{=} \sup_{U,V}\left\{\langle Y, UV^\top\rangle : \sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2 = \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2 \le 1\right\} \\
&= \sup_{\substack{a\in\mathbb{R}^n_+ : \|a\|_{\mathcal{R}}\le 1 \\ b\in\mathbb{R}^m_+ : \|b\|_{\mathcal{C}}\le 1}} \sup_{U,V}\left\{\langle Y, UV^\top\rangle : \left\|U_{(i)}\right\|_2 \le a_i \;\forall i,\; \left\|V_{(j)}\right\|_2 \le b_j \;\forall j\right\} \\
&\le K_G \cdot \sup_{\substack{a\in\mathbb{R}^n_+ : \|a\|_{\mathcal{R}}\le 1 \\ b\in\mathbb{R}^m_+ : \|b\|_{\mathcal{C}}\le 1}} \sup_{U,V}\left\{\langle Y, uv^\top\rangle : |u_i| \le a_i \;\forall i,\; |v_j| \le b_j \;\forall j\right\} \\
&= K_G \cdot \sup_{u,v}\left\{\langle Y, uv^\top\rangle : \|u\|_{\mathcal{R}} \le 1, \|v\|_{\mathcal{C}} \le 1\right\} \\
&= K_G \cdot \sup_{X}\left\{\langle Y, X\rangle : X \in \mathrm{Conv}\left\{uv^\top : u \in \mathbb{R}^n, v \in \mathbb{R}^m, \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\}\right\} \\
&= \sup_{X}\left\{\langle Y, X\rangle : X \in K_G \cdot \mathrm{Conv}\left\{uv^\top : u \in \mathbb{R}^n, v \in \mathbb{R}^m, \|u\|_{\mathcal{R}} = \|v\|_{\mathcal{C}} = 1\right\}\right\} .
\end{aligned}$$

As in [1], this is sufficient to prove the result. Above, the step marked (*) is true because, given any U and V with

$$\frac{1}{2}\left(\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2\right) \le 1 ,$$

we can replace U and V with $U' := U\cdot\omega$ and $V' := V\cdot\omega^{-1}$, where $\omega := \sqrt[4]{\frac{\sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2}{\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2}}$. This will give $U'V'^\top = UV^\top$, and

$$\begin{aligned}
\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U'_{(i)}\right\|_2^2 = \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V'_{(j)}\right\|_2^2 &= \sqrt{\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2 \cdot \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2} \\
&\le \frac{1}{2}\left(\sup_{\mathbf{r}\in\mathcal{R}}\sum_i \mathbf{r}_i \left\|U_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}}\sum_j \mathbf{c}_j \left\|V_{(j)}\right\|_2^2\right) \le 1 .
\end{aligned}$$

C Proof of Theorem 3

Following the strategy of Srebro & Shraibman (2005), we will use the Rademacher complexity to bound this excess risk. By Theorem 8 of Bartlett & Mendelson (2002)[1], we know that

$$\mathbb{E}_S\left[\sum_{ij}\mathbf{p}_{ij}\left|Y_{ij}-\widehat{X}_{ij}\right| - \inf_{\|X\|_{(\mathcal{R},\mathcal{C})}\leq\sqrt{k}}\sum_{ij}\mathbf{p}_{ij}\left|Y_{ij}-X_{ij}\right|\right]$$
$$= \mathcal{O}\left(\mathbb{E}_S\left[\widehat{\mathcal{R}}_S\left(\left\{X\in\mathbb{R}^{n\times m}:\|X\|_{(\mathcal{R},\mathcal{C})}\leq\sqrt{k}\right\}\right)\right]\right) , \quad (5)$$

where the expected Rademacher complexity is defined as

$$\mathbb{E}_S\left[\widehat{\mathcal{R}}_S\left(\left\{X\in\mathbb{R}^{n\times m}:\|X\|_{(\mathcal{R},\mathcal{C})}\leq\sqrt{k}\right\}\right)\right] := \frac{1}{s}\mathbb{E}_{S,\nu}\left[\sup_{\|X\|_{(\mathcal{R},\mathcal{C})}\leq\sqrt{k}}\sum_t \nu_t\cdot X_{i_tj_t}\right] ,$$

where $\nu\in\{\pm1\}^s$ is a random vector of independent unbiased signs, generated independently from S.

Now we bound the Rademacher complexity. By scaling, it is sufficient to consider the case $k=1$. The main idea for this proof is to first show that, for any X with $\|X\|_{(\mathcal{R},\mathcal{C})}\leq 1$, we can decompose X into a sum $X'+X''$ where $\|X'\|_{\max}\leq K_G$ and $\|X''\|_{\mathrm{tr}(\widetilde{\mathbf{p}})}\leq 2K_G\gamma^{-1/2}$, where $\widetilde{\mathbf{p}}$ represents the smoothed row and column marginals with smoothing parameter $\zeta = 1/2$, and where $K_G\leq 1.79$ is Grothendieck's constant. We will then use known Rademacher complexity bounds for the classes of matrices that have bounded max norm and bounded smoothed weighted trace norm.

To construct the decomposition of X, we start with a vector decomposition lemma, proved below.

Lemma 1. *Suppose $\mathcal{R}\supseteq\mathcal{R}^{\times}_{1/2,\gamma}$. Then for any $u\in\mathbb{R}^n$ with $\|u\|_{\mathcal{R}}=1$, we can decompose u into a sum $u=u'+u''$ such that $\|u'\|_\infty\leq 1$ and $\|u''\|_{\widetilde{\mathbf{p}}_{\mathrm{row}}} := \sum_i\widetilde{\mathbf{p}}_{i\bullet}u_i''^2\leq\gamma^{-1/2}$.*

Next, by Theorem 2, we can write

$$X = K_G\cdot\sum_{l=1}^{\infty}t_l\cdot u_lv_l^\top ,$$

where $t_l\geq 0$, $\sum_{l=1}^\infty t_l = 1$, and $\|u_l\|_{\mathcal{R}}=\|v_l\|_{\mathcal{C}}=1$ for all l. Applying Lemma 1 to u_l and to v_l for each l, we can write $u_l = u_l'+u_l''$ and $v_l=v_l'+v_l''$, where

$$\|u_l'\|_\infty\leq 1,\ \|u_l''\|_{\widetilde{\mathbf{p}}_{\mathrm{row}}}\leq\gamma^{-1/2},\ \|v_l'\|_\infty\leq 1,\ \|v_l''\|_{\widetilde{\mathbf{p}}_{\mathrm{col}}}\leq\gamma^{-1/2} .$$

Then

$$X = K_G\cdot\left(\sum_{l=1}^\infty t_l\cdot u_l'v_l'^\top + \sum_{l=1}^\infty t_l\cdot u_l'v_l''^\top + \sum_{l=1}^\infty t_l\cdot u_l''v_l^\top\right) =: K_G\left(X_1+X_2+X_3\right) .$$

Furthermore, $\|u_l'\|_{\widetilde{\mathbf{p}}_{\mathrm{row}}}\leq\|u_l'\|_\infty\leq 1$, and $\|v_l\|_{\widetilde{\mathbf{p}}_{\mathrm{row}}}\leq\|v_l\|_{\mathcal{C}}\leq 1$. Applying Srebro and Shraibman [1]'s convex hull bounds for the trace norm and max norm (stated in Section 4 of the main paper), we see that $\|X_1\|_{\max}\leq 1$, and that that $\|X_i\|_{\mathrm{tr}(\widetilde{\mathbf{p}})}\leq\gamma^{-1/2}$ for $i=2,3$. Defining $X'=X_1$ and $X''=X_2+X_3$, we have the desired decomposition.

Applying this result to every X in the class $\left\{X\in\mathbb{R}^{n\times m}:\|X\|_{(\mathcal{R},\mathcal{C})}\leq 1\right\}$, we see that

$$\mathbb{E}_S\left[\widehat{\mathcal{R}}_S\left(\left\{X\in\mathbb{R}^{n\times m}:\|X\|_{(\mathcal{R},\mathcal{C})}\leq 1\right\}\right)\right]$$
$$\leq \mathbb{E}_S\left[\widehat{\mathcal{R}}_S\left(\{X':\|X'\|_{\max}\leq K_G\}\right)\right] + \mathbb{E}_S\left[\widehat{\mathcal{R}}_S\left(\left\{X'':\|X''\|_{\mathrm{tr}(\widetilde{\mathbf{p}})}\leq K_G\cdot 2\gamma^{-1/2}\right\}\right)\right]$$
$$\leq K_G\cdot\mathcal{O}\left(\sqrt{\frac{n}{s}}\right) + K_G\cdot 2\gamma^{-1/2}\cdot\mathcal{O}\left(\sqrt{\frac{n\log(n)}{s}}+\frac{n\log(n)}{s}\right) ,$$

[1] The statement of their theorem gives a result that holds with high probability, but in the proof of this result they derive a bound in expectation, which we use here.

where the last step uses bounds on the Rademacher complexity of the max norm and weighted trace norm unit balls, shown in Theorem 5 of [1] and Theorem 3 of [2], respectively. Finally, we want to deal with the last term, $\frac{n\log(n)}{s}$, that is outside the square root. Since $s \geq n$ by assumption, we have $\frac{n\log(n)}{s} \leq \sqrt{\frac{n\log^2(n)}{s}}$, and if $s \geq n\log(n)$, then we can improve this to $\frac{n\log(n)}{s} \leq \sqrt{\frac{n\log(n)}{s}}$. Returning to (5) and plugging in our bound on the Rademacher complexity, this proves the desired bound on the excess risk.

C.1 Proof of Lemma 1

For $u \in \mathbb{R}^n$ with $\|u\|_{\mathcal{R}} = 1$, we need to find a decomposition $u = u' + u''$ such that $\|u'\|_\infty \leq 1$ and $\|u''\|_{\widetilde{\mathbf{p}}_{\mathrm{row}}} = \sqrt{\sum_i \widetilde{\mathbf{p}}_{i\bullet} u_i''^2} \leq \gamma^{-1/2}$. Without loss of generality, assume $|u_1| \geq \cdots \geq |u_n|$. Find $N \in \{1, \dots, n\}$ and $t \in (0, 1]$ so that $\sum_{i=1}^{N-1} \widetilde{\mathbf{p}}_{i\bullet} + t \cdot \widetilde{\mathbf{p}}_{N\bullet} = \gamma^{-1}$, and let

$$\mathbf{r} = \gamma \cdot (\widetilde{\mathbf{p}}_{1\bullet}, \dots, \widetilde{\mathbf{p}}_{(N-1)\bullet}, t \cdot \widetilde{\mathbf{p}}_{N\bullet}, 0, \dots, 0) \in \Delta_{[n]} \ .$$

Clearly, $\mathbf{r}_i \leq \gamma \cdot \widetilde{\mathbf{p}}_{i\bullet}$ for all i, and so $\mathbf{r} \in \mathcal{R}^\times_{1/2,\gamma} \subseteq \mathcal{R}$.

Now let $u'' = (u_1, \dots, u_{N-1}, \sqrt{t} \cdot u_N, 0, \dots, 0)$, and set $u' = u - u''$. We then calculate

$$\|u''\|^2_{\widetilde{\mathbf{p}}_{\mathrm{row}}} = \sum_{i=1}^{N-1} \widetilde{\mathbf{p}}_{i\bullet} u_i^2 + t \cdot \widetilde{\mathbf{p}}_{N\bullet} u_N^2 = \gamma^{-1} \sum_{i=1}^{n} \mathbf{r}_i u_i^2 \leq \gamma^{-1} \|u\|^2_{\mathcal{R}} \leq \gamma^{-1} \ .$$

Finally, we want to show that $\|u'\|_\infty \leq 1$. Since $u'_i = 0$ for $i < N$, we only need to bound $|u'_i|$ for each $i \geq N$. We have

$$1 = \|u\|^2_{\mathcal{R}} \geq \sum_{i'=1}^{n} \mathbf{r}_{i'} u_{i'}^2 \geq \sum_{i'=1}^{N} \mathbf{r}_{i'} u_{i'}^2 \overset{(*)}{\geq} u_i^2 \cdot \sum_{i'=1}^{N} \mathbf{r}_{i'} \overset{(\#)}{=} u_i^2 \geq u_i'^2 \ ,$$

where the step marked (*) uses the fact that $|u_{i'}| \geq |u_i|$ for all $i' \leq N$, and the step marked (#) comes from the fact that $\mathbf{r}$ is supported on $\{1, \dots, N\}$. This is sufficient.

D Proof of Proposition 1

Let $L_0 = \mathrm{Loss}(\widehat{X})$. Then, by definition,

$$\widehat{X} = \arg\min \left\{ \mathrm{Penalty}_{(\beta,\tau)}(X) : \mathrm{Loss}(X) \leq L_0 \right\} \ .$$

Then to prove the lemma, it is sufficient to show that for some $t \in [0, 1]$,

$$\widehat{X} = \arg\min \left\{ \|X\|_{(\mathcal{R}_{(t)}, \mathcal{C}_{(t)})} : \mathrm{Loss}(X) \leq L_0 \right\} \ ,$$

where we set

$$\mathcal{R}_{(t)} = \left\{ \mathbf{r} \in \Delta_{[n]} : \mathbf{r}_i \geq \frac{t}{1 + (n-1)t} \ \forall i \right\}, \ \mathcal{C}_{(t)} = \left\{ \mathbf{c} \in \Delta_{[m]} : \mathbf{c}_j \geq \frac{t}{1 + (m-1)t} \ \forall j \right\} \ .$$

Trivially, we can rephrase these definitions as

$$\mathcal{R}_{(t)} = \left\{ \frac{t}{1 + (n-1) \cdot t} \cdot (1, \dots, 1) + \frac{1-t}{1 + (n-1) \cdot t} \cdot \mathbf{r} : \mathbf{r} \in \Delta_{[n]} \right\} \text{ and}$$
$$\mathcal{C}_{(t)} = \left\{ \frac{t}{1 + (m-1) \cdot t} \cdot (1, \dots, 1) + \frac{1-t}{1 + (m-1) \cdot t} \cdot \mathbf{c} : \mathbf{c} \in \Delta_{[m]} \right\} \ . \quad (6)$$

Note that for any vectors $u \in \mathbb{R}^n_+$ and $v \in \mathbb{R}^m_+$,

$$\sup_{\mathbf{r} \in \Delta_{[n]}} \sum_i \mathbf{r}_i u_i = \max_i u_i \text{ and } \sup_{\mathbf{c} \in \Delta_{[m]}} \sum_j \mathbf{c}_j v_j = \max_j v_j \ . \quad (7)$$

Applying the SDP formulation of the local max norm (proved in Lemma 2 below), we have

$$\|X\|_{(\mathcal{R}_{(t)},\mathcal{C}_{(t)})} = \frac{1}{2}\inf\left\{\sup_{\mathbf{r}\in\mathcal{R}_{(t)}}\sum_i \mathbf{r}_i U_{ii} + \sup_{\mathbf{c}\in\mathcal{C}_{(t)}}\sum_j \mathbf{c}_j V_{jj} : \begin{pmatrix} U & X \\ X^\top & V \end{pmatrix} \succeq 0\right\}$$

$$\overset{\text{By (6) and (7)}}{=} \frac{1}{2}\inf\left\{\frac{t}{1+(n-1)\cdot t}\cdot\sum_i U_{ii} + \frac{1-t}{1+(n-1)\cdot t}\max_i U_{ii}\right.$$

$$\left. + \frac{t}{1+(m-1)\cdot t}\cdot\sum_j V_{jj} + \frac{1-t}{1+(m-1)\cdot t}\max_j V_{jj} : \begin{pmatrix} U & X \\ X^\top & V \end{pmatrix} \succeq 0\right\}$$

$$= \frac{\omega_t}{2}\inf\left\{t\sum_i A_{ii} + (1-t)\max_i A_{ii} + t\sum_j B_{jj} + (1-t)\max_j B_{jj} : \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0\right\}$$

$$= \frac{\omega_t}{2}\inf\left\{(1-t)\cdot \mathrm{M}(A,B) + t\cdot \mathrm{T}(A,B) : X \in \mathcal{X}_{A,B}\right\}, \quad (8)$$

where for the next-to-last step, we define

$$A = U\cdot\sqrt{\frac{1+(m-1)\cdot t}{1+(n-1)\cdot t}},\ B = V\cdot\sqrt{\frac{1+(n-1)\cdot t}{1+(m-1)\cdot t}},\ \omega_t = \frac{1}{\sqrt{(1+(n-1)\cdot t)(1+(m-1)\cdot t)}}\ ,$$

and for the last step, we define

$$\mathrm{T}(A,B) = \mathrm{trace}(A) + \mathrm{trace}(B),\ \mathrm{M}(A,B) = \max_i A_{ii} + \max_j B_{jj}\ ,$$

and

$$\mathcal{X}_{A,B} = \left\{X : \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0\right\}\ .$$

Next, we compare this to the (β,τ) penalty formulated in our main paper. Recall

$$\mathrm{Penalty}_{(\beta,\tau)}(X) = \inf_{X=AB^\top}\left\{\sqrt{\max_i \left\|A_{(i)}\right\|_2^2 + \max_j \left\|B_{(j)}\right\|_2^2}\cdot\sqrt{\sum_i \left\|A_{(i)}\right\|_2^2 + \sum_j \left\|B_{(j)}\right\|_2^2}\right\}\ .$$

Applying Lemma 3 below, we can obtain an equivalent SDP formulation of the penalty

$$\mathrm{Penalty}_{(\beta,\tau)}(X) = \inf_{A,B}\left\{\sqrt{\mathrm{M}(A,B)}\cdot\sqrt{\mathrm{T}(A,B)}\ :\ X \in \mathcal{X}_{A,B}\right\}\ . \quad (9)$$

Since $\mathrm{M}(A,B) \le \mathrm{T}(A,B) \le \max\{n,m\}\mathrm{M}(A,B)$, and since for any $x,y>0$ we know $\sqrt{xy} \le \frac{1}{2}\left(\alpha\cdot x + \alpha^{-1}\cdot y\right)$ for any $\alpha > 0$ with equality attained when $\alpha = \sqrt{y/x}$, we see that

$$\mathrm{Penalty}_{(\beta,\tau)}(\widehat{X}) = \frac{1}{2}\inf_{A,B}\left\{\inf_{\alpha\in[1,\sqrt{\max\{n,m\}}]}\left\{\alpha\cdot\mathrm{M}(A,B) + \alpha^{-1}\cdot\mathrm{T}(A,B)\right\}\ :\ \widehat{X}\in\mathcal{X}_{A,B}\right\}$$

$$= \inf_{\alpha\in[1,\sqrt{\max\{n,m\}}]}\left[\frac{1}{2}\inf_{A,B}\left\{\alpha\cdot\mathrm{M}(A,B) + \alpha^{-1}\cdot\mathrm{T}(A,B)\ :\ \widehat{X}\in\mathcal{X}_{A,B}\right\}\right]\ .$$

Since the quantity inside the square brackets is nonnegative and is continuous in α, and we are minimizing over α in a compact set, the infimum is attained at some $\widehat{\alpha}$, so we can write

$$\mathrm{Penalty}_{(\beta,\tau)}(\widehat{X}) = \frac{1}{2}\inf_{A,B}\left\{\widehat{\alpha}\cdot\mathrm{M}(A,B) + \widehat{\alpha}^{-1}\cdot\mathrm{T}(A,B)\ :\ \widehat{X}\in\mathcal{X}_{A,B}\right\}\ .$$

Recall that $\widehat{X}$ minimizes $\mathrm{Penalty}_{(\beta,\tau)}(X)$ subject to the constraint $\mathrm{Loss}(X) \leq L_0$. Setting $t := \frac{\widehat{\alpha}^{-1}}{\widehat{\alpha}+\widehat{\alpha}^{-1}}$, we get

$$\begin{aligned}
\widehat{X} &\in \arg\min_X \left\{ \inf_{A,B} \left\{ \widehat{\alpha} \cdot \mathrm{M}(A,B) + \widehat{\alpha}^{-1} \cdot \mathrm{T}(A,B) \;:\; X \in \mathcal{X}_{A,B} \right\} : \mathrm{Loss}(X) \leq L_0 \right\} \\
&= \arg\min_X \left\{ \inf_{A,B} \left\{ \frac{\widehat{\alpha}}{\widehat{\alpha}+\widehat{\alpha}^{-1}} \cdot \mathrm{M}(A,B) + \frac{\widehat{\alpha}^{-1}}{\widehat{\alpha}+\widehat{\alpha}^{-1}} \cdot \mathrm{T}(A,B) \;:\; X \in \mathcal{X}_{A,B} \right\} : \mathrm{Loss}(X) \leq L_0 \right\} \\
&= \arg\min_X \left\{ \inf_{A,B} \left\{ (1-t) \cdot \mathrm{M}(A,B) + t \cdot \mathrm{T}(A,B) \;:\; X \in \mathcal{X}_{A,B} \right\} : \mathrm{Loss}(X) \leq L_0 \right\} \\
&= \arg\min_X \left\{ \|X\|_{(\mathcal{R}_{(t)},\mathcal{C}_{(t)})} : \mathrm{Loss}(X) \leq L_0 \right\} ,
\end{aligned}$$

as desired.

E Computing the local max norm with an SDP

Lemma 2. *Suppose $\mathcal{R}$ and $\mathcal{C}$ are convex, and are defined by SDP-representable constraints. Then the $(\mathcal{R},\mathcal{C})$-norm can be calculated with the semidefinite program*

$$\|X\|_{(\mathcal{R},\mathcal{C})} = \frac{1}{2} \inf \left\{ \sup_{\mathbf{r}\in\mathcal{R}} \sum_i \mathbf{r}_i A_{ii} + \sup_{\mathbf{c}\in\mathcal{C}} \sum_j \mathbf{c}_j B_{jj} : \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0 \right\} .$$

In the special case where $\mathcal{R}$ and $\mathcal{C}$ are defined as in (8) in the main paper, then the norm is given by

$$\begin{aligned}
\|X\|_{(\mathcal{R},\mathcal{C})} = \frac{1}{2} \inf \Big\{ & a + R^\top a_1 + b + C^\top b_1 \;:\; a_{1i} \geq 0 \text{ and } a + a_{1i} \geq A_{ii}\ \forall i, \\
& b_{1j} \geq 0 \text{ and } b + b_{1j} \geq B_{jj}\ \forall j, \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0 \Big\} .
\end{aligned}$$

Proof. For the general case, based on Theorem 1 in the main paper, we only need to show that

$$\begin{aligned}
&\inf \left\{ \sup_{\mathbf{r}\in\mathcal{R}} \sum_i \mathbf{r}_i A_{ii} + \sup_{\mathbf{c}\in\mathcal{C}} \sum_j \mathbf{c}_j B_{jj} : \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0 \right\} \\
&\qquad = \inf \left(\sup_{\mathbf{r}\in\mathcal{R}} \sum_i \mathbf{r}_i \left\|A_{(i)}\right\|_2^2 + \sup_{\mathbf{c}\in\mathcal{C}} \sum_j \mathbf{c}_j \left\|B_{(j)}\right\|_2^2 : AB^\top = X \right) .
\end{aligned}$$

This is proved in Lemma 3 below.

For the special case where $\mathcal{R}$ and $\mathcal{C}$ are defined by element-wise bounds, we return to the proof of Theorem 1 given in Section A, where we see that

$$2\|X\|_{(\mathcal{R},\mathcal{C})} = \inf_{\substack{AB^\top = X, a,b\in\mathbb{R} \\ a_1\in\mathbb{R}^n_+, b_1\in\mathbb{R}^m_+}} \left\{ a + R^\top a_1 + b + C^\top b_1 \;:\; a + a_{1i} \geq \left\|A_{(i)}\right\|_2^2\ \forall i,\ b + b_{1j} \geq \left\|B_{(j)}\right\|_2^2\ \forall j \right\} .$$

Noting that $\left\|A_{(i)}\right\|_2^2 = (AA^\top)_{ii}$ and $\left\|B_{(j)}\right\|_2^2 = (BB^\top)_{jj}$, we again use Lemma 3 to see that this is equivalent to the SDP

$$\begin{aligned}
\inf \Big\{ & a + R^\top a_1 + b + C^\top b_1 \;:\; a_{1i} \geq 0 \text{ and } a + a_{1i} \geq A_{ii}\ \forall i, \\
& b_{1j} \geq 0 \text{ and } b + b_{1j} \geq B_{jj}\ \forall j, \begin{pmatrix} A & X \\ X^\top & B \end{pmatrix} \succeq 0 \Big\} .
\end{aligned}$$

$\square$

Lemma 3. *Let $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be any function that is nondecreasing in each coordinate and let $X \in \mathbb{R}^{n\times m}$ be any matrix. Then*

$$\inf\left\{f\left(\left\|A_{(1)}\right\|_2^2,\ldots,\left\|A_{(n)}\right\|_2^2,\left\|B_{(1)}\right\|_2^2,\ldots,\left\|B_{(m)}\right\|_2^2\right) : AB^\top = X\right\}$$
$$= \inf\left\{f\left(\Phi_{11},\ldots,\Phi_{nn},\Psi_{11},\ldots,\Psi_{mm}\right) : \begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0\right\}\ ,$$

where the factorization $AB^\top = X$ is assumed to be of arbitrary dimension, that is, $A \in \mathbb{R}^{n\times k}$ and $B \in \mathbb{R}^{m\times k}$ for arbitrary $k \in \mathbb{N}$.

Proof. We follow similar arguments as in Lemma 14 in [3], where this equality is shown for the special case of calculating a trace norm.

For convenience, we write

$$g(A,B) = f\left(\left\|A_{(1)}\right\|_2^2,\ldots,\left\|A_{(n)}\right\|_2^2,\left\|B_{(1)}\right\|_2^2,\ldots,\left\|B_{(m)}\right\|_2^2\right)$$

and

$$h(\Phi,\Psi) = f\left(\Phi_{11},\ldots,\Phi_{nn},\Psi_{11},\ldots,\Psi_{mm}\right)\ .$$

Then we would like to show that

$$\inf\left\{g(A,B) : AB^\top = X\right\} = \inf\left\{h(\Phi,\Psi) : \begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0\right\}\ .$$

First, take any factorization $AB^\top = X$. Let $\Phi = AA^\top$ and $\Psi = BB^\top$. Then $\begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0$, and we have $g(A,B) = h(\Phi,\Psi)$ by definition. Therefore,

$$\inf\left\{g(A,B) : AB^\top = X\right\} \geq \inf\left\{h(\Phi,\Psi) : \begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0\right\}\ .$$

Next, take any Φ and Ψ such that $\begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0$. Take a Cholesky decomposition

$$\begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} = \begin{pmatrix}A & 0\\ B & C\end{pmatrix} \cdot \begin{pmatrix}A & 0\\ B & C\end{pmatrix}^\top = \begin{pmatrix}AA^\top & AB^\top\\ BA^\top & BB^\top + CC^\top\end{pmatrix}\ .$$

From this, we see that $AB^\top = X$, that $\Phi_{ii} = \left\|A_{(i)}\right\|_2^2$ for all i, and that $\Psi_{jj} \geq \left\|B_{(j)}\right\|_2^2$ for all j. Since f is nondecreasing in each coordinate, we have $h(\Phi,\Psi) \geq g(A,B)$. Therefore, we see that

$$\inf\left\{g(A,B) : AB^\top = X\right\} \leq \inf\left\{h(\Phi,\Psi) : \begin{pmatrix}\Phi & X\\ X^\top & \Psi\end{pmatrix} \succeq 0\right\}\ .$$

□

References

[1] N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. *18th Annual Conference on Learning Theory (COLT)*, pages 545–560, 2005.

[2] R. Foygel, R. Salakhutdinov, O. Shamir, and N. Srebro. Learning with the weighted trace-norm under arbitrary sampling distributions. *Advances in Neural Information Processing Systems*, 24, 2011.

[3] N. Srebro. *Learning with matrix factorizations*. PhD thesis, Citeseer, 2004.