Learning with the Weighted Trace-norm under Arbitrary Sampling Distributions

Rina Foygel
Department of Statistics
University of Chicago
rina@uchicago.edu

Ruslan Salakhutdinov
Department of Statistics
University of Toronto
rsalakhu@ustat.toronto.edu

Ohad Shamir
Microsoft Research New England
ohadsh@microsoft.com

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

Supplementary Materials

A Proofs for the i.i.d. sampling setting

A.1 Proof of Theorem 1

We first fill in the details for the Rademacher bound in the case that p has uniform row- and column-marginals. Define

$$Q_t = \sigma_t \frac{e_{i_t,j_t}}{\sqrt{p^r(i_t)\,p^c(j_t)}} \in \mathbb{R}^{n\times m}\ .$$

We need to calculate R and ρ^2 such that $\|Q_t\|_{\mathrm{sp}} \le R$ (almost surely) and

$$\rho^2 = \max\left\{ \left\| \sum \mathbf{E}\left[Q_t^T Q_t\right] \right\|_{\mathrm{sp}}, \left\| \sum \mathbf{E}\left[Q_t Q_t^T\right] \right\|_{\mathrm{sp}} \right\}.$$

For each t, Q_t is just a matrix with a single non-zero entry of magnitude $\frac{1}{\sqrt{p^r(i)p^c(j)}}$, for some i, j, and so $\|Q_t\|_{\mathrm{sp}} \le \max_{ij} \frac{1}{\sqrt{p^r(i)p^c(j)}} \doteq R$.

The matrix $Q_t Q_t^T \in \mathbb{R}^{n\times n}$ is equal to $\frac{e_{i,i}}{p^r(i)p^c(j)}$ with probability $p(i,j)$. Hence $\mathbf{E}\left[Q_t^T Q_t\right]$ is a diagonal matrix with entries $\sum_j \frac{p(i,j)}{p^r(i)p^c(j)}$. Similar arguments apply to $Q_t^T Q_t$. Multiplying by s, and recalling the spectral norm of a diagonal matrix is simply the maximal magnitude element, we have:

$$\rho^2 = s \cdot \max\left\{ \max_i \sum_j \frac{p(i,j)}{p^r(i)\,p^c(j)}, \max_j \sum_i \frac{p(i,j)}{p^r(i)\,p^c(j)} \right\}\ .$$

This completes the proof for the case that p has uniform row- and column- marginals.

Next we turn to the case that p is a product distribution, $p = p^r \times p^c$ (with possibly non-uniform marginals). For any $X \in \mathcal{W}_r[p]$, define

$$Z(X) = \left(X_{ij} \mathbb{I}\left\{ p(i,j) \ge \frac{\log(n)}{s\sqrt{nm}} \right\} \right)_{ij}\ .$$

Let $\mathcal{Z} = \{Z(X) : X \in \mathcal{W}_r[p]\}$.

We can then follow the proof of the square-root bound in Theorem 1, with a modified definition of Q_t:

$$Q_t = \sigma_t \frac{e_{i_t,j_t} \mathbb{I}\left\{p\left(i_t,j_t\right) \geq \frac{\log(n)}{s\sqrt{nm}}\right\}}{\sqrt{p^r\left(i_t\right)p^c\left(j_t\right)}} \ .$$

Proceeding as in the proof for Theorem 1, we obtain $R \leq \sqrt{\frac{s\sqrt{nm}}{\log(n)}}$ and $\rho^2 \leq sn$, and thus

$$\mathbf{E}_{S\sim p}\left[\hat{\mathcal{R}}_S(\mathcal{Z})\right] = \mathbf{O}\left(\sqrt{\frac{rn\log(n)}{s}}\right) \ .$$

Therefore, by [1],

$$\mathbf{E}\left[\sup_{X\in\mathcal{W}_r[p]} L_p(Z(X)) - \hat{L}_S(Z(X))\right] \leq \mathbf{O}\left(l\cdot\sqrt{\frac{rn\log(n)}{s}}\right) \ ,$$
$$\mathbf{E}\left[\sup_{X\in\mathcal{W}_r[p]} \hat{L}_S(Z(X)) - L_p(Z(X))\right] \leq \mathbf{O}\left(l\cdot\sqrt{\frac{rn\log(n)}{s}}\right) \ .$$

Next, let $I = \left(\sqrt{p\left(i,j\right)}\mathbb{I}\left\{p\left(i,j\right) < \frac{\log(n)}{s\sqrt{nm}}\right\}\right)_{ij}$. For any matrix M, define

$$\left\|M\right\|_{F(p^r,p^c)} = \left\|\mathrm{diag}\left(p^r\right)^{1/2} M \mathrm{diag}\left(p^c\right)^{1/2}\right\|_F \ .$$

Now take any M with $\left\|M\right\|_{F(p^r,p^c)} \leq 1$. Let $M' = \mathrm{diag}\left(p^r\right)^{1/2} M \mathrm{diag}\left(p^c\right)^{1/2}$, then $\left\|M'\right\|_F \leq 1$. We have

$$\sum_{ij:p(i,j)<\frac{\log(n)}{s\sqrt{nm}}} p\left(i,j\right) M_{ij} = \sum_{ij} I_{ij} M'_{ij} = \left\langle I, M'\right\rangle \leq \left\|I\right\|_F \cdot \left\|M'\right\|_F$$
$$\leq \left\|I\right\|_F = \sqrt{\sum_{ij} p\left(i,j\right)\mathbb{I}\left\{p\left(i,j\right) < \frac{\log(n)}{s\sqrt{nm}}\right\}} \leq \sqrt{nm\cdot\frac{\log(n)}{s\sqrt{nm}}} = \sqrt{\frac{\sqrt{nm}\log(n)}{s}} \ .$$

Since $\left\|M\right\|_F \leq \left\|M\right\|_{\mathrm{tr}}$ for any matrix M, we then have, for any $X \in \mathcal{W}_r\left[p\right]$, $\left\|X\right\|_{F(p^r,p^c)} \leq \left\|X\right\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}$, and so

$$\left|L_p(X) - L_p(Z(X))\right| = \left|\sum_{ij\notin\mathcal{I}} p\left(i,j\right)\left(\ell(X_{ij},Y_{ij}) - \ell(0,Y_{ij})\right)\right| \leq l\cdot\sum_{ij\notin\mathcal{I}} p\left(i,j\right)\left|X_{ij}\right| \leq \sqrt{\frac{l^2 r\sqrt{nm}\log(n)}{s}} \ .$$

And, fixing some $X^* \in \mathcal{W}_r[p]$ such that $L_p(X^*) = \inf_{X \in \mathcal{W}_r[p]} L_p(X)$,

$$\begin{aligned}
&\mathbf{E}\left[\sup_{X \in \mathcal{W}_r[p]} \hat{L}_S(Z(X)) - \hat{L}_S(X)\right] + \mathbf{E}\left[\hat{L}_S(X^*) - \hat{L}_S(Z(X^*))\right] \\
&= \mathbf{E}\left[\sup_{X \in \mathcal{W}_r[p]} \frac{1}{s}\sum_{t=1}^{s} \mathbb{I}\left\{(i_t, j_t) \notin \mathcal{I}\right\}\left(\ell(0, Y_{i_t j_t}) - \ell(X_{i_t j_t}, Y_{i_t j_t})\right)\right] \\
&\quad + \mathbf{E}\left[\frac{1}{s}\sum_{t=1}^{s} \mathbb{I}\left\{(i_t, j_t) \notin \mathcal{I}\right\}\left(\ell(X^*_{i_t j_t}, Y_{i_t j_t}) - \ell(0, Y_{i_t j_t})\right)\right] \\
&= \mathbf{E}\left[\sup_{X \in \mathcal{W}_r[p]} \frac{1}{s}\sum_{t=1}^{s} \mathbb{I}\left\{(i_t, j_t) \notin \mathcal{I}\right\}\left(\ell(X^*_{i_t j_t}, Y_{i_t j_t}) - \ell(X_{i_t j_t}, Y_{i_t j_t})\right)\right] \\
&\leq \mathbf{E}\left[\sup_{X \in \mathcal{W}_r[p]} \frac{1}{s}\sum_{t=1}^{s} \mathbb{I}\left\{(i_t, j_t) \notin \mathcal{I}\right\}\ell(X^*_{i_t j_t}, Y_{i_t j_t})\right] \leq l \cdot \mathbf{E}\left[\frac{1}{s}\sum_{t=1}^{s} \mathbb{I}\left\{(i_t, j_t) \notin \mathcal{I}\right\}|X^*_{i_t j_t}|\right] \\
&= l \cdot \mathbf{E}\left[\mathbb{I}\left\{(i_1, j_1) \notin \mathcal{I}\right\}|X^*_{i_1 j_1}|\right] = l \cdot \sum_{ij \notin \mathcal{I}} p(i,j)\,|X^*_{ij}| \leq \sqrt{\frac{l^2 r\sqrt{nm}\log(n)}{s}}
\end{aligned}$$

Then writing

$$L_p(\hat{X}_S) - L_p(X^*) = (L_p(\hat{X}_S) - L_p(Z(\hat{X}_S))) + (L_p(Z(\hat{X}_S)) - \hat{L}_S(Z(\hat{X}_S))) + (\hat{L}_S(Z(\hat{X}_S)) - \hat{L}_S(\hat{X}_S))$$
$$+ (\hat{L}_S(\hat{X}_S) - \hat{L}_S(X^*)) + (\hat{L}_S(X^*) - \hat{L}_S(Z(X^*))) + (\hat{L}_S(Z(X^*)) - L_p(Z(X^*))) + (L_p(Z(X^*)) - L_p(X^*)) ,$$

we obtain

$$\mathbf{E}\left[L_p(\hat{X}_S) - L_p(X^*)\right] \leq \mathbf{O}\left(\sqrt{\frac{l^2 rn\log(n)}{s}}\right) .$$

A.2 Proof of Theorem 2

Assume ℓ is l-Lipschitz and b-bounded, and $r \geq 1$. We will show that (for any p)

$$\mathbf{E}_{S\sim p}\left[\hat{\mathcal{R}}_S(\ell \circ \mathcal{W}_r[p])\right] = \mathbf{O}\left((l+b)\cdot\sqrt[3]{\frac{rn\log(n)}{s}}\right) .$$

Given a sample S, define

$$T_S^0 = \left\{t \,:\, p^r(i_t) \text{ or } p^c(j_t) < \sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\}, T_S^1 = \{1,\ldots,s\}\backslash T_S^0 .$$

We have

$$\begin{aligned}
&\hat{\mathcal{R}}_S(\ell \circ \mathcal{W}_r[p]) = \mathbf{E}_{\sigma\sim\{\pm1\}^s}\left[\sup_{\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}} \frac{1}{s}\sum_{t=1}^{s} \sigma_t \cdot \ell(X_{i_t j_t}, Y_{i_t j_t})\right] \\
&\leq \mathbf{E}_\sigma\left[\sup_{\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}} \frac{1}{s}\sum_{t\in T_S^0} \sigma_t \cdot \ell(X_{i_t j_t}, Y_{i_t j_t})\right] + \mathbf{E}_\sigma\left[\sup_{\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}} \frac{1}{s}\sum_{t\in T_S^1} \sigma_t \cdot \ell(X_{i_t j_t}, Y_{i_t j_t})\right]
\end{aligned}$$

Bounding the first term,

$$\mathbf{E}_\sigma\left[\sup_{\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}} \frac{1}{s}\sum_{t\in T_S^0} \sigma_t \cdot \ell(X_{i_t j_t}, Y_{i_t j_t})\right] \leq \mathbf{E}_\sigma\left[\frac{1}{s}\sum_{t\in T_S^0} |\sigma_t| \cdot b\right] = \frac{b}{s}\cdot\left|T_S^0\right| .$$

In expectation over S,

$$\begin{aligned}
&\mathbf{E}_S\left[\frac{b}{s}\cdot\left|T_S^0\right|\right] = b\cdot\mathbf{E}_{ij\sim p}\left[\mathbb{I}\left\{p^r\left(i\right)\ \text{or}\ p^c\left(j\right) < \sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\}\right] \\
&= b\cdot\sum_{ij} p\left(i,j\right)\mathbb{I}\left\{p^r\left(i\right)\ \text{or}\ p^c\left(j\right) < \sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\} \\
&\leq \left[b\cdot\sum_{i:p^r(i)<\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}}\sum_j p\left(i,j\right)\right] + \left[b\cdot\sum_{j:p^c(j)<\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}}\sum_i p\left(i,j\right)\right] \\
&= \left[b\cdot\sum_{i:p^r(i)<\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}} p^r\left(i\right)\right] + \left[b\cdot\sum_{j:p^c(j)<\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}} p^c\left(j\right)\right] \\
&\leq bn\cdot\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}} + bm\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}} \leq 2\sqrt[3]{\frac{l^2 brn\log(n)}{s}}\ .
\end{aligned}$$

To bound the second term, we use the fact that $\left\|\text{abs}(X)\right\|_{\text{tr}} \leq \left\|X\right\|_{\text{tr}}$ for any matrix X, where $\text{abs}(X)$ is the matrix defined via $\text{abs}(X)_{ij} = |X_{ij}|$. We have

$$\begin{aligned}
&\mathbf{E}_\sigma\left[\sup_{\left\|X\right\|_{\text{tr}(p^r,p^c)}\leq\sqrt{r}}\frac{1}{s}\sum_{t\in T_S^1}\sigma_t\cdot\ell(X_{i_t j_t}, Y_{i_t j_t})\right] \leq l\cdot\mathbf{E}_\sigma\left[\sup_{\left\|X\right\|_{\text{tr}(p^r,p^c)}\leq\sqrt{r}}\frac{1}{s}\sum_{t\in T_S^1}\sigma_t\cdot|X_{i_t j_t}|\right] \\
&= l\cdot\mathbf{E}_\sigma\left[\sup_{\left\|X'\right\|_{\text{tr}}\leq\sqrt{r}}\frac{1}{s}\sum_{t\in T_S^1}\frac{\sigma_t}{\sqrt{p^r\left(i_t\right)p^c\left(j_t\right)}}\cdot|X'_{i_t j_t}|\right] \leq l\cdot\mathbf{E}_\sigma\left[\sup_{\left\|X''\right\|_{\text{tr}}\leq\sqrt{r}}\frac{1}{s}\sum_{t\in T_S^1}\frac{\sigma_t}{\sqrt{p^r\left(i_t\right)p^c\left(j_t\right)}}\cdot X''_{i_t j_t}\right] \\
&= l\sqrt{r}\cdot\mathbf{E}_\sigma\left[\left\|\frac{1}{s}\sum_{t=1}^{s}\sigma_t\frac{e_{(i_t,j_t)}\mathbb{I}\left\{p^r\left(i_t\right),p^c\left(j_t\right)\geq\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\}}{\sqrt{p^r\left(i_t\right)p^c\left(j_t\right)}}\right\|_{\text{sp}}\right],
\end{aligned}$$

Defining $Q_t = \sigma_t\frac{e_{(i_t,j_t)}\mathbb{I}\left\{p^r(i_t),p^c(j_t)\geq\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\}}{\sqrt{p^r(i_t)p^c(j_t)}}$, we can follow identical arguments as in the proof of the first bound of this theorem. We have

$$\left\|Q_t\right\|_{\text{sp}} \leq \max_{ij}\frac{\mathbb{I}\left\{p^r\left(i\right),p^c\left(j\right)\geq\sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\}}{\sqrt{p^r\left(i\right)p^c\left(j\right)}} \leq \sqrt[3]{\frac{b^2 sn^2}{l^2 r\log(n)}} \doteq R\ ,$$

and

$$\begin{aligned}
\rho^2 &\doteq \max\left\{\left\|\sum \mathbf{E}\left[Q_t^T Q_t\right]\right\|_{\mathrm{sp}}, \left\|\sum \mathbf{E}\left[Q_t Q_t^T\right]\right\|_{\mathrm{sp}}\right\} \\
&\leq s \cdot \max\left\{\max_i \sum_j \frac{p\left(i,j\right)\mathbb{I}\left\{p^r\left(i\right), p^c\left(j\right) \geq \sqrt[3]{\frac{l^2 r\log(n)}{b^2 s n^2}}\right\}}{p^r\left(i\right)p^c\left(j\right)}, \right. \\
&\qquad\qquad \left. \max_j \sum_i \frac{p\left(i,j\right)\mathbb{I}\left\{p^r\left(i\right), p^c\left(j\right) \geq \sqrt[3]{\frac{l^2 r\log(n)}{b^2 s n^2}}\right\}}{p^r\left(i\right)p^c\left(j\right)}\right\} \\
&\leq s \cdot \sqrt[3]{\frac{b^2 s n^2}{l^2 r\log(n)}} \cdot \max\left\{\max_i \sum_j \frac{p\left(i,j\right)}{p^r\left(i\right)}, \max_j \sum_i \frac{p\left(i,j\right)}{p^r\left(i\right)}\right\} = \sqrt[3]{\frac{b^2 s^4 n^2}{l^2 r\log(n)}}
\end{aligned}$$

Then applying [2], we get

$$\begin{aligned}
\mathbf{E}_\sigma\left[\sup_{\|X\|_{\mathrm{tr}(p^r,p^c)} \leq \sqrt{r}} \frac{1}{s}\sum_{t\in T_S^1} \sigma_t \cdot \ell(X_{i_t j_t}, Y_{i_t j_t})\right] &= \frac{l\sqrt{r}}{s}\mathbf{E}_{S,\sigma}\left[\left\|\sum_{t=1}^{s} Q_t\right\|_{\mathrm{sp}}\right] \\
&\leq \mathbf{O}\left(\frac{l\sqrt{r}}{s}\left(\rho\sqrt{\log(n)} + R\log(n)\right)\right) \\
&\leq \mathbf{O}\left(\frac{l\sqrt{r}}{s}\left(\sqrt[6]{\frac{b^2 s^4 n^2}{l^2 r\log(n)}}\sqrt{\log(n)} + \sqrt[3]{\frac{b^2 s n^2}{l^2 r\log(n)}}\log(n)\right)\right) \\
&\leq \mathbf{O}\left(l^{2/3} b^{1/3}\sqrt[3]{\frac{rn\log(n)}{s}} + l^{1/3} b^{2/3}\left(\sqrt[3]{\frac{rn\log(n)}{s}}\right)^2\right) .
\end{aligned}$$

If $s \geq rn\log(n)$, then this proves the bound. If not, then the result is trivial, since $L_p(X) \leq b$ for any X.

A.3 Proof of Theorem 4

Throughout this section, assume $s \geq 24n\log(n)$. (If this is not the case, then we only need to prove excess error $\leq \mathbf{O}(l\sqrt{r})$, which is trivial given the class $\mathcal{W}_r\left[\check{p}\right]$.) We also assume $s \leq \mathbf{O}\left(nm\log(nm)\right)$. (If this is not the case, then with high probability, we observe all entries of the matrix and obtain optimal recovery.) The lemmas which are cited in this proof, are proved below.

Define

$$X^* = \arg\min_{X\in\mathcal{W}_r[\check{p}]} L_p(X), \quad r^* = \|X^*\|^2_{\mathrm{tr}(p^r,p^c)} \leq r \ .$$

For any sample S, define

$$c(S) = \max\left\{0, \left\|\frac{1}{\sqrt{r^*}}X^*\right\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} - 1\right\} \ .$$

Then, for a fixed S,

$$\|(1-c(S))X^*\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} = \sqrt{r^*}(1-c(S))\left\|\frac{1}{\sqrt{r^*}}X^*\right\|_{\mathrm{tr}(\tilde{p}^r,\tilde{p}^c)} \leq \sqrt{r} \ \Rightarrow \ (1-c(S))X^* \in \mathcal{W}_r\left[\check{p}\right] \ .$$

Applying Lemma 1 and Theorem 3,

$$\begin{aligned}
&\mathbf{E}\left[L_p(\hat{X}_S) - \hat{L}_S(\hat{X}_S)\right] \leq \mathbf{E}\left[\sup_{X\in\mathcal{W}_r[\check{p}]}\left(L_p(X) - \hat{L}_S(X)\right)\right] \\
&\leq \mathbf{E}\left[\sup_{X\in 2\cdot\mathcal{W}_r[\tilde{p}]}\left(L_p(X) - \hat{L}_S(X)\right)\right] + \frac{8\sqrt{l^2rnm}}{n^2} \leq \mathbf{O}\left(\sqrt{\frac{l^2rn\log(n)}{s}}\right) + \frac{8\sqrt{l^2rnm}}{n^2} \\
&\leq \mathbf{O}\left(\sqrt{\frac{l^2rn\log(n)}{s}}\right) .
\end{aligned}$$

And, similarly,

$$\begin{aligned}
&\mathbf{E}\left[\hat{L}_S((1-c(S))X^*) - L_p((1-c(S))X^*)\right] \leq \mathbf{E}\left[\sup_{X\in\mathcal{W}_r[\check{p}]}\left(\hat{L}_S(X) - L_p(X)\right)\right] \\
&\leq \mathbf{E}\left[\sup_{X\in 2\cdot\mathcal{W}_r[\tilde{p}]}\left(\hat{L}_S(X) - L_p(X)\right)\right] + \frac{8\sqrt{l^2rnm}}{n^2} \leq \mathbf{O}\left(\sqrt{\frac{l^2rn\log(n)}{s}}\right) + \frac{8\sqrt{l^2rnm}}{n^2} \\
&\leq \mathbf{O}\left(\sqrt{\frac{l^2rn\log(n)}{s}}\right) .
\end{aligned}$$

By definition, since $(1-c(S))X^* \in \mathcal{W}_r[\check{p}]$,

$$\mathbf{E}\left[\hat{L}_S(\hat{X}_S) - \hat{L}_S((1-c(S))X^*)\right] \leq 0 .$$

Finally, by Lemma 3,

$$\mathbf{E}\left[L_p((1-c(S))X^*) - L_p(X^*)\right] \leq \sqrt{\frac{2l^2rn}{s}} .$$

Combining all of the above, we get

$$\mathbf{E}\left[L_p(\hat{X}_S) - L_p(X^*)\right] \leq O\left(\sqrt{\frac{l^2rn\log(n)}{s}}\right) .$$

A.3.1 Lemmas for Theorem 4

Lemma 1.

$$\begin{aligned}
&\mathbf{E}\left[\sup_{X\in\mathcal{W}_r[\check{p}]}\left(L_p(X) - \hat{L}_S(X)\right)\right] \leq \mathbf{E}\left[\sup_{X\in 2\cdot\mathcal{W}_r[\tilde{p}]}\left(L_p(X) - \hat{L}_S(X)\right)\right] + \frac{8\sqrt{l^2rnm}}{n^2} . \\
&\mathbf{E}\left[\sup_{X\in\mathcal{W}_r[\check{p}]}\left(\hat{L}_S(X) - L_p(X)\right)\right] \leq \mathbf{E}\left[\sup_{X\in 2\cdot\mathcal{W}_r[\tilde{p}]}\left(\hat{L}_S(X) - L_p(X)\right)\right] + \frac{8\sqrt{l^2rnm}}{n^2} .
\end{aligned}$$

Proof. By Lemma 2, with probability at least $1-2n^{-2}$, for all i, j,

$$\check{p}^r(i) \geq \frac{1}{2}\tilde{p}^r(i) , \; \check{p}^c(j) \geq \frac{1}{2}\tilde{p}^c(j) .$$

Let A be the event that these inequalities hold. If A occurs, then for any $X \in \mathcal{W}_r[\check{p}]$,

$$\begin{aligned}
\|X\|_{\mathrm{tr}(p^r,p^c)} &= \left\|\mathrm{diag}\left(\tilde{p}^r(i)\right)^{1/2} X \mathrm{diag}\left(\tilde{p}^c(j)\right)^{1/2}\right\|_{\mathrm{tr}} \\
&= \left\|\mathrm{diag}\left(\frac{\tilde{p}^r(i)}{\check{p}^r(i)}\right)^{1/2} \mathrm{diag}\left(\check{p}^r(i)\right)^{1/2} X \mathrm{diag}\left(\check{p}^c(j)\right)^{1/2} \mathrm{diag}\left(\frac{\tilde{p}^c(j)}{\check{p}^c(j)}\right)^{1/2}\right\|_{\mathrm{tr}} \\
&\leq 2\left\|\mathrm{diag}\left(\check{p}^r(i)\right)^{1/2} X \mathrm{diag}\left(\check{p}^c(j)\right)^{1/2}\right\|_{\mathrm{tr}} = 2\|X\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} \leq 2\sqrt{r} .
\end{aligned}$$

In this case, $\mathcal{W}_r\left[\check{p}\right] \subset 2 \cdot \mathcal{W}_r\left[\tilde{p}\right]$, and therefore,

$$\sup_{X \in \mathcal{W}_r[\check{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \\ \leq \sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] .$$

Next we consider the case that A does not occur. For any $X \in \mathcal{W}_r\left[\check{p}\right]$,

$$|X|_\infty \leq \|X\|_F = 2\sqrt{nm} \left\| \mathrm{diag}\left(\frac{1}{2n}\mathbf{1}_n\right)^{1/2} X \mathrm{diag}\left(\frac{1}{2m}\mathbf{1}_m\right)^{1/2} \right\|_F \\ \leq 2\sqrt{nm} \left\| \mathrm{diag}\left(\frac{1}{2n}\mathbf{1}_n\right)^{1/2} X \mathrm{diag}\left(\frac{1}{2m}\mathbf{1}_m\right)^{1/2} \right\|_{\mathrm{tr}} \leq 2\sqrt{nm}\, \|X\|_{\mathrm{tr}(\check{p}^r, \check{p}^c)} \leq 2\sqrt{rnm} .$$

Therefore,

$$\sup_{X \in \mathcal{W}_r[\check{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \\ \leq \sup_{X \in \mathcal{W}_r[\check{p}]} \left[\sum_{ij} p\left(i,j\right) \left(\ell(X_{ij}, Y_{ij}) - \ell(0, Y_{ij})\right) - \frac{1}{s}\sum_t \left(\ell(X_{i_t j_t}, Y_{i_t j_t}) - \ell(0, Y_{i_t j_t})\right)\right] \\ \leq l \cdot \sup_{X \in \mathcal{W}_r[\check{p}]} \left[\sum_{ij} p\left(i,j\right) \cdot |X_{ij}| + \frac{1}{s}\sum_t |X_{i_t j_t}|\right] \\ \leq l \cdot \sup_{X \in \mathcal{W}_r[\check{p}]} \left[\sum_{ij} p\left(i,j\right) \cdot 2\sqrt{rnm} + \frac{1}{s}\sum_t 2\sqrt{rnm}\right] \leq 4\sqrt{l^2 rnm}$$

And so,

$$\mathbf{E}_S \left[\sup_{X \in \mathcal{W}_r[\check{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right]\right] \\ = \mathbf{E}_S \left[\sup_{X \in \mathcal{W}_r[\check{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \cdot \mathbb{I}\left\{A\right\}\right] \\ + \mathbf{E}_S \left[\sup_{X \in \mathcal{W}_r[\check{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \cdot \mathbb{I}\left\{A^c\right\}\right] \\ \leq \mathbf{E}_S \left[\sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \cdot \mathbb{I}\left\{A\right\}\right] + P\left(A^c\right) \cdot 4\sqrt{l^2 rnm} \\ \leq \mathbf{E}_S \left[\sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \cdot \mathbb{I}\left\{A\right\}\right] + \frac{8\sqrt{l^2 rnm}}{n^2} \\ \leq \mathbf{E}_S \left[\sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right]\right] + \frac{8\sqrt{l^2 rnm}}{n^2} .$$

where the last step is true because, since $\mathbf{0}_{n\times m} \in 2 \cdot \mathcal{W}_r\left[\tilde{p}\right]$, for any S,

$$\sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left[\left(L_p(X) - L_p(\mathbf{0}_{n\times m})\right) - \left(\hat{L}_S(X) - \hat{L}_S(\mathbf{0}_{n\times m})\right)\right] \geq 0 .$$

And, $\mathbf{E}_S\left[L_p(\mathbf{0}_{n\times m}) - \hat{L}_S(\mathbf{0}_{n\times m})\right] = 0$, so therefore,

$$\mathbf{E}_S \left[\sup_{X \in \mathcal{W}_r[\check{p}]} \left(L_p(X) - \hat{L}_S(X)\right)\right] \leq \mathbf{E} \left[\sup_{X \in 2\cdot\mathcal{W}_r[\tilde{p}]} \left(L_p(X) - \hat{L}_S(X)\right)\right] + \frac{8\sqrt{l^2 rnm}}{n^2} .$$

The second claim can be proved with identical arguments. □

Lemma 2. *With probability at least* $1-2n^{-2}$*, for all* i *and all* j*,*

$$\check{p}^r(i) \geq \frac{1}{2}\tilde{p}^r(i)\,,\ \check{p}^c(j) \geq \frac{1}{2}\tilde{p}^c(j)\ .$$

Proof. Take any row i. Suppose that $p^r(i) \leq \frac{1}{n}$. Then $\tilde{p}^r(i) \leq \frac{1}{n}$, while $\check{p}^r(i) = \frac{1}{2}\left(\hat{p}^r(i) + \frac{1}{n}\right) \geq \frac{1}{2n}$. Therefore, in this case, $\check{p}^r(i) \geq \frac{1}{2}\tilde{p}^r(i)$ with probability 1.

Next, suppose that $p^r(i) > \frac{1}{n}$. Then, by the Chernoff inequality,

$$P\left(\hat{p}^r(i) < \frac{1}{2}p^r(i)\right) = P\left(\mathrm{Bin}(s, p^r(i)) < sp^r(i)\left(1-\frac{1}{2}\right)\right) \leq e^{-\frac{sp^r(i)}{8}}$$
$$\leq e^{-\frac{s}{8n}} \leq e^{-3\log(n)} = n^{-3}\ .$$

Therefore, with probability at least $1-n^{-3}$, $\hat{p}^r(i) \geq \frac{1}{2}p^r(i)$, and so

$$\check{p}^r(i) = \frac{1}{2}\left(\hat{p}^r(i) + \frac{1}{n}\right) \geq \frac{1}{2}\left(\frac{1}{2}p^r(i) + \frac{1}{n}\right) \geq \frac{1}{2}\tilde{p}^r(i)\ .$$

Therefore, for any row i, with probability at least $1-n^{-3}$, $\check{p}^r(i) \geq \frac{1}{2}\tilde{p}^r(i)$. The same reasoning applies to every column j. Therefore, with probability at least $1-2n^{-2}$, the statement holds for all i and all j. □

Lemma 3. *Fix* X^* *with* $\|X^*\|^2_{\mathrm{tr}(\tilde{p}^r,\tilde{p}^c)} = r^* \leq r$*, and define*

$$c(S) = \max\left\{0, \left\|\frac{1}{\sqrt{r^*}}X^*\right\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} - 1\right\}\ .$$

Then

$$\mathbf{E}\left[L_p((1-c(S))X^*) - L_p(X^*)\right] \leq \sqrt{\frac{2l^2rn}{s}}\ .$$

Proof.

$$L_p((1-c(S))X^*) - L_p(X^*) = \sum_{ij} p(i,j)\left(\ell((1-c(S))X^*_{ij}, Y_{ij}) - \ell(X^*_{ij}, Y_{ij})\right)$$
$$\leq l\cdot\sum_{ij} p(i,j)\left|(1-c(S))X^*_{ij} - X^*_{ij}\right| = l\cdot c(S)\cdot\sum_{ij} p(i,j)\left|X^*_{ij}\right|$$
$$= l\cdot c(S)\cdot\sum_{ij}\frac{p(i,j)}{\sqrt{\tilde{p}^r(i)\,\tilde{p}^c(j)}}\cdot\sqrt{\tilde{p}^r(i)\,\tilde{p}^c(j)}\cdot\left|X^*_{ij}\right|$$

Defining $M = \left(\frac{p(i,j)}{\sqrt{\tilde{p}^r(i)\tilde{p}^c(j)}}\right)_{ij}$,

$$= l\cdot c(S)\cdot\left\langle M, \left(\mathrm{diag}\left(\tilde{p}^r(i)\right)^{1/2} X^* \mathrm{diag}\left(\tilde{p}^c(j)\right)^{1/2}\right)_{ij}\right\rangle$$
$$\leq l\cdot c(S)\cdot\|M\|_{\mathrm{sp}}\cdot\left\|\left(\mathrm{diag}\left(\tilde{p}^r(i)\right)^{1/2} X^* \mathrm{diag}\left(\tilde{p}^c(j)\right)^{1/2}\right)_{ij}\right\|_{\mathrm{tr}}$$
$$\leq l\sqrt{r}\cdot c(S)\cdot\|M\|_{\mathrm{sp}}\ .$$

Now we show that $\|M\|_{\mathrm{sp}} \leq 2$. Take any unit vectors $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$. Then

$$u^T M v = \sum_{ij} p(i,j)\cdot\sqrt{\frac{u_i^2}{\tilde{p}^r(i)}}\cdot\sqrt{\frac{v_j^2}{\tilde{p}^c(j)}} \leq \frac{1}{2}\sum_{ij} p(i,j)\left(\frac{u_i^2}{\tilde{p}^r(i)} + \frac{v_j^2}{\tilde{p}^c(j)}\right)$$
$$= \frac{1}{2}\sum_i p^r(i)\cdot\frac{u_i^2}{\tilde{p}^r(i)} + \frac{1}{2}\sum_j p^c(j)\cdot\frac{v_j^2}{\tilde{p}^c(j)} \leq \frac{1}{2}\sum_i 2u_i^2 + \frac{1}{2}\sum_j 2v_j^2 = 2\,.$$

So, by Lemma 4,

$$\mathbf{E}\left[L_p((1-c(S))X^*) - L_p(X^*)\right] \leq 2l\sqrt{r}\cdot\mathbf{E}\left[c(S)\right] \leq 2l\sqrt{r}\cdot\sqrt{\frac{n}{2s}}\,.$$

□

Lemma 4. *For any p, for any fixed X with $\|X\|_{\mathrm{tr}(\tilde{p}^r,\tilde{p}^c)} = 1$,*

$$\mathbf{E}\left[\max\{0, \|X\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} - 1\}\right] \leq \sqrt{\frac{n}{2s}}\,.$$

Proof. By properties of the trace-norm [3], we can write $\mathrm{diag}\left(\tilde{p}^r\right)^{1/2} X \mathrm{diag}\left(\tilde{p}^c\right)^{1/2} = AB^T$, where $\|A\|_F^2 = \|B\|_F^2 = \|X\|_{\mathrm{tr}(p^r,p^c)} = 1$. Define

$$D_1 = \mathrm{diag}\left(\check{p}^r\right)\mathrm{diag}\left(\tilde{p}^r\right)^{-1},\ D_2 = \mathrm{diag}\left(\check{p}^c\right)\mathrm{diag}\left(\tilde{p}^c\right)^{-1}\ .$$

Then, by properties of the trace-norm [3],

$$\begin{aligned}
&\|X\|_{\mathrm{tr}(\check{p}^r,\check{p}^c)} = \left\|\mathrm{diag}\left(\check{p}^r\right)^{1/2} X \mathrm{diag}\left(\check{p}^c\right)^{1/2}\right\|_{\mathrm{tr}} = \left\|\left(D_1^{1/2}A\right)\left(D_2^{1/2}B\right)^T\right\|_{\mathrm{tr}} \\
&\leq \frac{1}{2}\left\|D_1^{1/2}A\right\|_F^2 + \frac{1}{2}\left\|D_2^{1/2}B\right\|_F^2 \\
&= \frac{1}{2}\sum_i \frac{\check{p}^r(i)}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{2}\sum_j \frac{\check{p}^c(j)}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2 \\
&= \frac{1}{4}\sum_i \frac{\hat{p}^r(i)}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\hat{p}^c(j)}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2 \\
&= \frac{1}{4}\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2\,,
\end{aligned}$$

where N_i^r is the number of samples in row i, and N_j^c is the number of samples in column j. Clearly,

$$\begin{aligned}
&\mathbf{E}\left[\frac{1}{4}\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2\right] \\
&= \frac{1}{4}\sum_i \frac{sp^r(i)}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{sp^c(j)}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2 \\
&= \frac{1}{2}\|A\|_F^2 + \frac{1}{2}\|B\|_F^2 = 1\,.
\end{aligned}$$

And, we can compute

$$\mathrm{Var}(N_i^r) \leq sp^r(i)\,,\ \mathrm{Cov}(N_i^r, N_{i'}^r) < 0,\ \mathrm{Var}(N_j^c) \leq sp^c(j)\,,\ \mathrm{Cov}(N_j^c, N_{j'}^c) < 0\,.$$

Therefore,

$$
\begin{aligned}
&\text{Var}\left(\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2\right) \\
&\leq 2\text{Var}\left(\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2\right) + 2\text{Var}\left(\sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2\right) \\
&= \sum_i \frac{1}{s^2\tilde{p}^r(i)^2}\text{Var}(N_i^r)\|A_{(i)}\|_2^4 + 2\sum_{i<i'} \frac{1}{s^2\tilde{p}^r(i)\,\tilde{p}^r(i')}\text{Cov}(N_i^r, N_{i'}^r)\|A_{(i)}\|_2^2\|A_{(i')}\|_2^2 \\
&\quad + \sum_j \frac{1}{s^2\tilde{p}^c(j)}\text{Var}(N_j^c)\|B_{(j)}\|_2^4 + 2\sum_{j<j'} \frac{1}{s^2\tilde{p}^c(j)\,\tilde{p}^c(j')}\text{Cov}(N_j^c, N_{j'}^c)\|B_{(j)}\|_2^2\|B_{(j')}\|_2^2 \\
&\leq \sum_i \frac{1}{s^2\tilde{p}^r(i)^2}\text{Var}(N_i^r)\|A_{(i)}\|_2^4 + \sum_j \frac{1}{s^2\tilde{p}^c(j)}\text{Var}(N_j^c)\|B_{(j)}\|_2^4 \\
&\leq \sum_i \frac{sp^r(i)}{s^2\tilde{p}^r(i)^2}\|A_{(i)}\|_2^4 + \sum_j \frac{sp^c(j)}{s^2\tilde{p}^c(j)}\|B_{(j)}\|_2^4
\end{aligned}
$$

Since $\tilde{p}^r(i) \geq \frac{1}{2}p^r(i)$ and $\tilde{p}^r(i) \geq \frac{1}{2n}$, and similarly for the columns,

$$
\begin{aligned}
&\leq \sum_i \frac{4n}{s}\|A_{(i)}\|_2^4 + \sum_j \frac{4m}{s}\|B_{(j)}\|_2^4 \leq \frac{4n}{s}\left(\sum_i \|A_{(i)}\|_2^2\right)^2 + \frac{4m}{s}\left(\sum_j \|B_{(j)}\|_2^2\right) \\
&\leq \frac{4n}{s}\|A\|_F^4 + \frac{4m}{s}\|B\|_F^4 \leq \frac{4(n+m)}{s}\,.
\end{aligned}
$$

So, we have

$$
\begin{aligned}
&\mathbf{E}\left[\max\{0, \|X\|_{\text{tr}(\check{p}^r,\check{p}^c)} - 1\}\right] \\
&\leq \mathbf{E}\left[\max\left\{0, \frac{1}{4}\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2 - 1\right\}\right] \\
&\leq \sqrt{\text{Var}\left(\frac{1}{4}\sum_i \frac{N_i^r}{s\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{N_j^c}{s\tilde{p}^c(j)}\|B_{(j)}\|_2^2 + \frac{1}{4}\sum_i \frac{\frac{1}{n}}{\tilde{p}^r(i)}\|A_{(i)}\|_2^2 + \frac{1}{4}\sum_j \frac{\frac{1}{n}}{\tilde{p}^c(j)}\|B_{(j)}\|_2^2\right)} \\
&\leq \sqrt{\frac{(n+m)}{4s}}
\end{aligned}
$$

□

B Proofs for the transductive setting

B.1 Proof of Theorem 5

Let $\overline{S} \subset [n] \times [m]$ be a subset of size $2s$. Let $\overline{p}$ denote the smoothed empirical marginals of $\overline{S}$.

Now choose any $S \subset \overline{S}$, a training set of size s. Without loss of generality, write $\overline{S} = \{(i_1, j_1), \ldots, (i_{2s}, j_{2s})\}$ and $S = \{(i_1, j_1), \ldots, (i_s, j_s)\}$.

First, we bound transductive Rademacher complexity. By Lemma 12 in [3], for any sample S,

$$\begin{aligned}
\hat{\mathcal{R}}_S(\mathcal{W}_r[\overline{p}]) &= \mathbf{E}_{\sigma\sim\{\pm1\}^s}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{t=1}^{s}\sigma_t X_{i_t j_t}\right] \\
&= \mathbf{E}_{\sigma\sim\{\pm1\}^s}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}X_{ij}\left(\sum_{t:(i_t,j_t)=(i,j)}\sigma_t\right)\right] \\
&\le \mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}X_{ij}\sigma_{ij}\cdot\#\{t:(i_t,j_t)=(i,j),1\le t\le s\}\right] \\
&= \mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}X_{ij}\sigma_{ij}\cdot\mathbb{I}\{(i,j)\in S\}\right].
\end{aligned}$$

Now define matrix Σ via

$$\Sigma_{ij} = \frac{\mathbb{I}\{(i,j)\in S\}}{s\sqrt{\overline{p}^r(i)\,\overline{p}^c(j)}}.$$

We have

$$\begin{aligned}
\mathbf{E}_{S\sim p}\left[\hat{\mathcal{R}}_S(\mathcal{W}_r[\overline{p}])\right] &\le \mathbf{E}_S\left[\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}X_{ij}\sigma_{ij}\cdot\mathbb{I}\{(i,j)\in S\}\right]\right] \\
&= \mathbf{E}_S\left[\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\sum_{ij}\left(\sqrt{p^r(i)}X_{ij}\sqrt{p^c(j)}\right)\sigma_{ij}\Sigma_{ij}\right]\right] \\
&= \mathbf{E}_S\left[\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X:\|X\|_{\mathrm{tr}}\le\sqrt{r}}X_{ij}\sigma_{ij}\Sigma_{ij}\right]\right] \\
&= \sqrt{r}\cdot\mathbf{E}_S\left[\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\|\sigma\bullet\Sigma\|_{\mathrm{sp}}\right]\right],
\end{aligned}$$

where $\sigma\bullet\Sigma$ is the element-wise product of Σ with the random sign matrix $\sigma=(\sigma_{ij})$. By [4],

$$\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\|\sigma\bullet\Sigma\|_{\mathrm{sp}}\right] \le \mathbf{O}\left(\log^{1/4}(n+m)\right)\cdot\max\left\{\max_i\left\|\Sigma_{(i)}\right\|_2,\max_j\left\|\Sigma^{(j)}\right\|_2\right\}.$$

We now bound $\left\|\Sigma_{(i)}\right\|_2$ and $\left\|\Sigma^{(j)}\right\|_2$. Fix any i. Then

$$\left\|\Sigma_{(i)}\right\|_2^2 = \sum_j\Sigma_{ij}^2 = \sum_{j=1}^{m}\frac{\mathbb{I}\{(i,j)\in S\}}{(s\overline{p}^r(i))\cdot(s\overline{p}^c(j))} \le \sum_{j=1}^{m}\frac{\mathbb{I}\left\{(i,j)\in\overline{S}\right\}}{(s\overline{p}^r(i))\cdot\left(s\cdot\frac{1}{2m}\right)}$$

$$\le \sum_{j=1}^{m}\frac{\#\{t:(i_t,j_t)=(i,j),1\le t\le 2s\}}{\left(\frac{1}{4}\#\left\{t:i_t=i,1\le t\le 2s\right\}\right)\cdot\left(s\cdot\frac{1}{2m}\right)} \le \frac{\#\{t:i_t=i,1\le t\le 2s\}}{\left(\frac{1}{4}\#\left\{t:i_t=i,1\le t\le 2s\right\}\right)\cdot\left(s\cdot\frac{1}{2m}\right)} \le \frac{8m}{s}.$$

Similarly, for all j, $\left\|\Sigma^{(j)}\right\|_2^2 \le \frac{8n}{s}$. Therefore,

$$\begin{aligned}
\mathbf{E}_{S\sim p}\left[\hat{\mathcal{R}}_S(\mathcal{W}_r[\overline{p}])\right] &\le \sqrt{r}\cdot\mathbf{E}_S\left[\mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\|\sigma\bullet\Sigma\|_{\mathrm{sp}}\right]\right] \\
&\le \sqrt{r}\cdot\mathbf{E}_S\left[\mathbf{O}\left(\log^{1/4}(n)\right)\cdot\max\left\{\max_i\left\|\Sigma_{(i)}\right\|_2,\max_j\left\|\Sigma^{(j)}\right\|_2\right\}\right] \\
&\le \mathbf{O}\left(\sqrt{\frac{rn\log^{1/2}(n)}{s}}\right).
\end{aligned}$$

Applying Theorem 5 of [5] (using integration to obtain a bound in expectation from a bound in probability),

$$\mathbf{E}_S\left[\hat{L}_{\overline{S}\setminus S}(\hat{X}_S) - \inf_{X\in\mathcal{W}_r[\overline{p}]}\hat{L}_{\overline{S}\setminus S}(X)\right] \le \mathbf{O}\left(\sqrt{\frac{l^2rn\log^{1/2}(n)+b^2}{s}}\right).$$

B.2 Transductive version of Theorem 1

Let $\overline{p}$ now denote the (unsmoothed) empirical marginals of $\overline{S}$. If $\overline{p}^r(i) \geq \frac{1}{Cn}$ and $\overline{p}^c(j) \geq \frac{1}{Cm}$ for all i, j, defining

$$\hat{X}_S = \arg\min_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_S(X) ,$$

we can then show that, for an l-Lipschitz loss ℓ bounded by b, in expectation over the split of $\overline{S}$ into training set S and test set T,

$$\hat{L}_T(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_T(X) + \mathbf{O}\left(C^{1/2} l \cdot \sqrt{\frac{rn\log^{1/2}(n) + b^2}{s}}\right) .$$

We prove this by following identical arguments as in the proof of Theorem 5. We define

$$\Sigma_{ij} = \frac{\mathbb{I}\{(i,j) \in S\}}{s\sqrt{\overline{p}^r(i)\,\overline{p}^c(j)}} ,$$

and obtain $\left\|\Sigma_{(i)}\right\|_2^2, \left\|\Sigma^{(j)}\right\|_2^2 \leq \frac{2Cn}{s}$ for all i, j, which yields

$$\mathbf{E}_S\left[\hat{L}_{\overline{S}\setminus S}(\hat{X}_S) - \inf_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_{\overline{S}\setminus S}(X)\right] \leq \mathbf{O}\left(\sqrt{\frac{Cl^2 rn\log^{1/2}(n) + b^2}{s}}\right) .$$

In fact, we can obtain the same result with a weaker requirement on $\overline{p}$, namely

$$\frac{s}{n}\max\left\{\max_i \|\Sigma_{(i)}\|_2^2, \max_j \|\Sigma^{(j)}\|_2^2\right\} \leq \max\left\{\max_i \frac{1}{m}\sum_{j=1}^{m} \frac{\frac{1}{s}\mathbb{I}\left\{(i,j) \in \overline{S}\right\}}{\overline{p}^r(i)\,\overline{p}^c(j)}, \max_j \frac{1}{n}\sum_{i=1}^{n} \frac{\frac{1}{s}\mathbb{I}\left\{(i,j) \in \overline{S}\right\}}{\overline{p}^r(i)\,\overline{p}^c(j)}\right\} \leq C .$$

For instance, this quantity is likely to be bounded if $\overline{S}$ is a sample drawn from a product distribution on the matrix.

B.3 Transductive version of Theorem 2

Let $\overline{p}$ now denote the (unsmoothed) empirical marginals of $\overline{S}$. We define

$$\hat{X}_S = \arg\min_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_S(X) ,$$

we can then show that, for an l-Lipschitz loss ℓ bounded by b, without any requirements on $\overline{p}$, in expectation over the split of $\overline{S}$ into training set S and test set T,

$$\hat{L}_T(\hat{X}_S) \leq \inf_{X \in \mathcal{W}_r[\overline{p}]} \hat{L}_T(X) + \mathbf{O}\left((l+b) \cdot \sqrt[3]{\frac{rn\log(n)}{s}}\right) .$$

We prove this by combining the proof techniques used in the proofs of Theorems 2 and 5. Define

$$T_S^0 = \left\{t \,:\, 1 \leq t \leq 2s, p^r(i_t) \text{ or } p^c(j_t) < \sqrt[3]{\frac{l^2 r\log(n)}{b^2 sn^2}}\right\} , T_S^1 = \{1, \ldots, 2s\}\setminus T_S^0 .$$

We then have

$$\hat{\mathcal{R}}_S(\ell \circ \mathcal{W}_r[\overline{p}]) = \mathbf{E}_{\sigma\sim\{\pm1\}^s}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{t=1}^{s}\sigma_t\ell(X_{i_tj_t},Y_{i_tj_t})\right]$$

$$\leq \mathbf{E}_{\sigma\sim\{\pm1\}^{n\times m}}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}\ell(X_{i_tj_t},Y_{i_tj_t})\sigma_{ij}\cdot\mathbb{I}\left\{(i,j)\in S\right\}\right]$$

$$\leq \mathbf{E}_{\sigma}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}\ell(X_{i_tj_t},Y_{i_tj_t})\sigma_{ij}\cdot\mathbb{I}\left\{(i,j)\in S,\text{ and }\overline{p}^r(i),\overline{p}^c(j)\geq\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}\right]$$

$$+\mathbf{E}_{\sigma}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}\ell(X_{i_tj_t},Y_{i_tj_t})\sigma_{ij}\cdot\mathbb{I}\left\{(i,j)\in S,\text{ and }\overline{p}^r(i)\text{ or }\overline{p}^c(j)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}\right]$$

$\doteq$ (Term 1) + (Term 2); .

Now define matrix Σ via

$$\Sigma_{ij}=\frac{\mathbb{I}\left\{(i,j)\in S,\text{ and }\overline{p}^r(i),\overline{p}^c(j)\geq\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}}{s\sqrt{\overline{p}^r(i)\,\overline{p}^c(j)}}\,.$$

Following the same arguments as in the proof of Theorem 5, we obtain for all i,j,

$$\|\Sigma_{(i)}\|_2^2,\|\Sigma^{(j)}\|_2^2\leq\frac{4}{s}\cdot\sqrt[3]{\frac{b^2sn^2}{l^2r\log(n)}}$$

Therefore, using the same arguments as in the proof of Theorem 2,

$$\text{(Term 1)}\leq l\sqrt{r}\mathbf{O}\left(\log^{1/4}(n)\sqrt{\frac{4}{s}\cdot\sqrt[3]{\frac{b^2sn^2}{l^2r\log(n)}}}\right)=\mathbf{O}\left(\sqrt[3]{\frac{l^2brn\log(n)}{s}}\right)\,.$$

Next we have

$$\text{(Term 2)}=\mathbf{E}_{\sigma}\left[\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}\ell(X_{i_tj_t},Y_{i_tj_t})\sigma_{ij}\cdot\mathbb{I}\left\{(i,j)\in S,\text{ and }\overline{p}^r(i)\text{ or }\overline{p}^c(j)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}\right]$$

$$\leq\sup_{X\in\mathcal{W}_r[\overline{p}]}\frac{1}{s}\sum_{ij}\ell(X_{i_tj_t},Y_{i_tj_t})\cdot\mathbb{I}\left\{(i,j)\in S,\text{ and }\overline{p}^r(i)\text{ or }\overline{p}^c(j)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}$$

$$\leq\frac{1}{s}\sum_{ij}b\cdot\mathbb{I}\left\{(i,j)\in\overline{S},\text{ and }\overline{p}^r(i)\text{ or }\overline{p}^c(j)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right\}$$

$$\leq\frac{1}{s}\sum_{i:\overline{p}^r(i)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}}\left(\sum_{j:(i,j)\in\overline{S}}b\right)+\frac{1}{s}\sum_{j:\overline{p}^c(j)<\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}}\left(\sum_{i:(i,j)\in\overline{S}}b\right)$$

$$\leq\frac{2n}{s}\left(b\cdot2s\cdot\sqrt[3]{\frac{l^2r\log(n)}{b^2sn^2}}\right)\leq\mathbf{O}\left(\sqrt[3]{\frac{l^2rbn\log(n)}{s}}\right)\,.$$

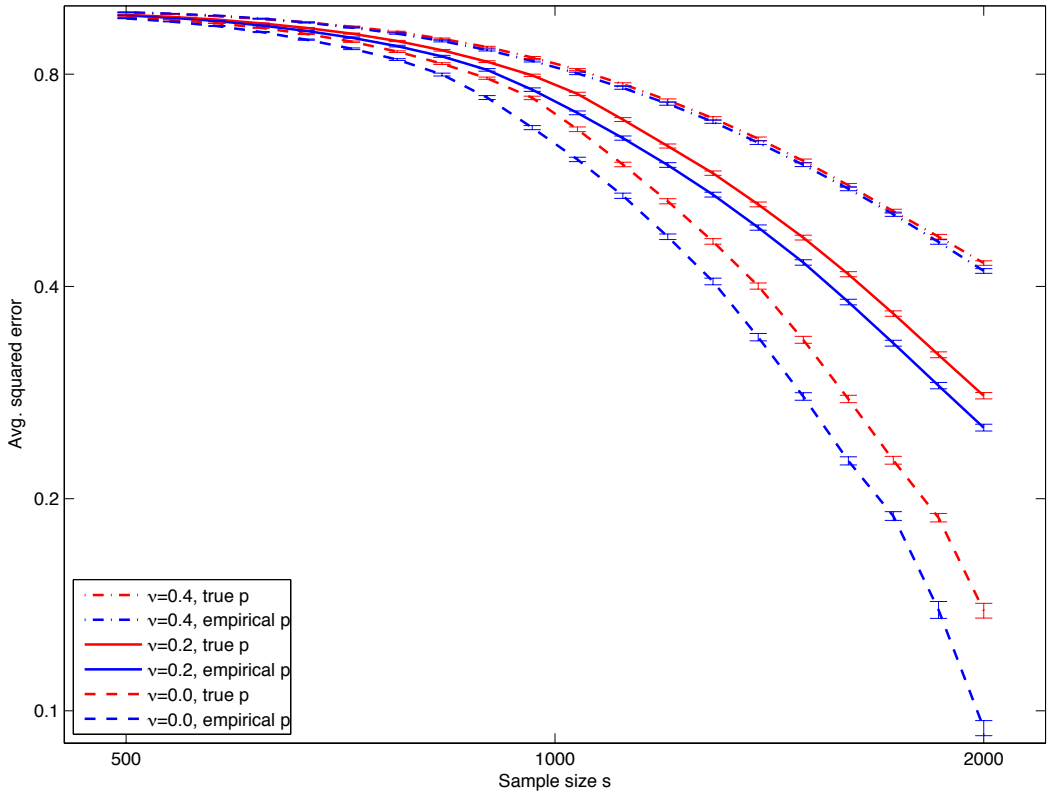
Combining the two, we get $\mathcal{R}_s(\ell\circ\mathcal{W}_r[\overline{p}])\leq\mathbf{O}\left(\sqrt[3]{\frac{l^2rbn\log(n)}{s}}\right)$, and therefore, in expectation over the split of $\overline{S}$ into S and T,

$$\hat{L}_T(\hat{X}_S)\leq\inf_{X\in\mathcal{W}_r[\overline{p}]}\hat{L}_T(X)+\mathbf{O}\left((l+b)\cdot\sqrt[3]{\frac{rn\log(n)}{s}}\right)\,.$$

C Simulations

C.1 Excess error comparison in the noiseless and noisy settings: larger figure

The figure below is a larger version of Figure 1(b) in the paper, with standard error bars added:



References

[1] P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

[2] J.A. Tropp. User-friendly tail bounds for sums of random matrices. *arXiv:1004.4389*, 2010.

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

[4] Y. Seginer. The expected norm of random matrices. *Combinatorics, Probability and Computing*, 9(2):149–166, 2000.

[5] O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. *24th Annual Conference on Learning Theory (COLT)*, 2011.