Convex Co-Embedding for Matrix Completion with Predictive Side Information Supplementary Material

1. Proof of Proposition 2

Proof: With the definitions of $\mathbf{h}$, $\tilde{\Lambda}$ $\tilde{Y}$ and $\tilde{\Omega}$, it is easy to see that the first two terms of the objective function in (9), denoted as $g(M)$, can be re-expressed as below

$$\begin{aligned} g(M) &= \|(\tilde{X}A - \overline{A})M + \mathbf{1}\mathbf{b}^\top\|_F^2 + \|\Omega \circ (\sqrt{\rho}\overline{A}M - \sqrt{\rho}Y)\|_F^2 \\ &= \|\mathbf{1}_{N,L} \circ \left((\tilde{X}A - \overline{A})M + \mathbf{1}\mathbf{b}^\top - \mathbf{0}_{N,L}\right)\|_F^2 + \\ &\quad \|\Omega \circ (\sqrt{\rho}\overline{A}M + \mathbf{0}\mathbf{b}^\top - \sqrt{\rho}Y)\|_F^2 \\ &= \|\tilde{\Omega} \circ (\tilde{\Lambda}M + \mathbf{h}\mathbf{b}^\top - \tilde{Y})\|_F^2 \end{aligned}$$

Hence the minimization problem (9) is equivalent to the following problem

$$\min_{M,\mathbf{b}} \|\tilde{\Omega} \circ (\tilde{\Lambda}M + \mathbf{h}\mathbf{b}^\top - \tilde{Y})\|_F^2 + \gamma\|M\|_{tr}$$

which is known to be equivalent to (11) by changing the nuclear norm regularizer into an inequality constraint with a proper τ value. ■

2. Proof of Theorem 1

Proof: The proof is given by presenting the following two lemmas. First, the relationship between the expected risk and the empirical risk can be built by applying the following lemma with Rademacher complexity.

Lemma 1 *(Chiang, Hsieh, and Dhillon 2015, Lemma1) Let $\mathcal{L}_\ell$ be a Lipschitz constant for the loss function ℓ with respect to its first argument, and assume it is bounded by $\mathcal{B}_\ell$. Let $\mathcal{R}(\mathcal{F}_\Theta)$ be the empirical Rademacher complexity of the function class $\mathcal{F}_\Theta$ defined as:*

$$\mathcal{R}(\mathcal{F}_\Theta) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}_\Theta} \frac{1}{m}\sum_{a=1}^{m} \sigma_a \ell(f(i_a, j_a), \tilde{Y}_{i_a j_a})\right],$$

where $\boldsymbol{\sigma} = \{\sigma_1, \cdots, \sigma_m\}$ are independent uniform $\{\pm 1\}$-valued random variables. Then for a constant $0 < \delta < 1$, with probability at least $1-\delta$ for all $f \in \mathcal{F}_\Theta$ we have the following bound on the expected risk

$$R_\ell(f) \le \widehat{R}_\ell(f) + 2\mathbb{E}_{\check{\Omega}}\left[\mathcal{R}(\mathcal{F}_\Theta)\right] + \mathcal{B}_\ell\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$$

It is clear that the expected risk is determined by both the empirical risk and the model complexity $\mathbb{E}_{\check{\Omega}}\left[\mathcal{R}(\mathcal{F}_\Theta)\right]$. The model complexity is related to the input features and the model structures, which are captured in $\{\tilde{\Lambda}, \mathbf{h}, \mathbf{b}\}$. It is also related to the constraint over M, which limits the space of the feasible function class. Assume $\|\mathbf{b}\|_2$ is upper bounded by $\mathcal{B}_b$, i.e., $\|\mathbf{b}\|_2 \le \mathcal{B}_b$. Below we provide a bound on the model complexity in terms of the properties of these related components.

Lemma 2 *Let $\kappa = \max\left(\sqrt{\rho}, \max_i\sqrt{\|\tilde{X}_i\|_2^2 + 1}\right)$, $q = \sqrt{N}\mathcal{B}_b$, $n_{\max} = \max(2N, L)$ and $d_{\max} = \max(d+N, L)$. Then the model complexity of the the function class $\mathcal{F}_\Theta$ is upper bounded by:*

$$\begin{aligned} &\mathbb{E}_{\check{\Omega}}\left[\mathcal{R}(\mathcal{F}_\Theta)\right] \le 2\tau\kappa\mathcal{L}_\ell\sqrt{\frac{\log 2d_{\max}}{m}} + \\ &\min\left\{2\mathcal{L}_\ell q\sqrt{\frac{\log 2n_{\max}}{m}}, \sqrt{9C\mathcal{L}_\ell\mathcal{B}_\ell\frac{q(\sqrt{2N}+\sqrt{L})}{m}}\right\} \end{aligned}$$

where C is a universal constant.

Based on our definitions of $\tilde{\Lambda}$ and $\mathbf{h}$, it is easy to verify that $\kappa = \max_i\|\tilde{\Lambda}_i\|_2$ and $\|\mathbf{h}\mathbf{b}^\top\|_{\text{tr}} \le q$. This Lemma can then be derived from the Lemma 2 of (Chiang, Hsieh, and Dhillon 2015) with its right side feature matrix as an identity matrix.

With the closed-form solution for $\mathbf{b}$ in (12), an upper bound for the Euclidean norm of $\mathbf{b}$ can be derived as following:

$$\begin{aligned} \|\mathbf{b}\|_2 &= \frac{1}{N}\left\|M^\top(\overline{A} - \tilde{X}A)^\top\mathbf{1}\right\|_2 \\ &= \frac{1}{N}\left\|\left[\mathbf{1}^\top\tilde{X}, -\mathbf{1}_{1,N}\right]M\right\|_2 \\ &\le \frac{1}{N}\left\|\left[\mathbf{1}^\top\tilde{X}, -\mathbf{1}_{1,N}\right]\right\|_2\|M\|_{sp} \\ &\le \frac{\|M\|_{tr}}{N}\sqrt{\|\mathbf{1}^\top\tilde{X}\|_2^2 + N} \\ &\le \tau\sqrt{\left\|(\mathbf{1}^\top\tilde{X})/N\right\|_2^2 + \frac{1}{N}} = \mathcal{B}_b^* \end{aligned}$$

Note if the feature matrix $\tilde{X}$ is already centered by its zero mean vector (a typical preprocessing step), i.e., $\mathbf{1}^\top\tilde{X} = \mathbf{0}^\top$, then we will have

$$\mathcal{B}_b^* = \tau\sqrt{\left\|(\mathbf{1}^\top\tilde{X})/N\right\|_2^2 + \frac{1}{N}} = \frac{\tau}{\sqrt{N}}$$

and

$$\|\mathbf{h}\mathbf{b}^\top\|_{tr} \le q = \sqrt{N}\mathcal{B}_b^* = \tau.$$

By combining Lemma 1, Lemma 2 and the q value derived above, we can get the upper bound for the expected risk of an optimal solution in Theorem 1.

3. Sample Complexity

The bound in Theorem 1 suggests a sample complexity of $O(\tau^2\log n_{\max})$. Below we provide a derivation for the upper bound of τ.

First by replacing $\mathbf{b}$ with the closed-form solution (12), we can re-express the empirical risk function in (11) as:

$$\|\check{\Omega} \circ (\check{\Lambda} M - \tilde{Y})\|_F^2,$$

where $\check{\Lambda} = [H\tilde{X}A - H\overline{A}; \sqrt{\rho}\overline{A}]$ and $H = I_N - \frac{1}{N}\mathbf{1}\mathbf{1}^\top$.

Now we construct a feasible solution while producing the constraint parameter τ simultaneously. Let $\mathcal{T}_\mu(\cdot)$ be a thresholding operator with parameter μ such that $\mathcal{T}_\mu(x) = xI_{[x \geq \mu]}$. Given the SVD of $\check{\Lambda}$ such as $\check{\Lambda} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where σ_i denotes the i-th largest singular value, we define $\check{\Lambda}_\mu = \sum_i \sigma_1 \mathcal{T}_\mu(\sigma_i/\sigma_1)\mathbf{u}_i\mathbf{v}_i^\top$. For $\mu \in (0, 1]$, we consider setting $\tau = \|\hat{M}\|_{tr}$ for a feasible solution $\hat{M}$:

$$\hat{M} = \arg\min_M \|\check{\Lambda}_\mu M - \tilde{Y}\|_F^2 = (\check{\Lambda}_\mu^\top \check{\Lambda}_\mu)^{-1}\check{\Lambda}_\mu^\top \tilde{Y} \quad (1)$$

We then have the following lemma, which shows that the nuclear norm of $\hat{M}$ is upper bounded by $O(\sqrt{n_{\max}})$:

Lemma 3 Given $\mu \in (0, 1]$, $\check{\kappa} = \max_i \|\check{\Lambda}_i\|_2$ and $\lambda = \frac{\min_i \|\check{\Lambda}_i\|_2}{\check{\kappa}}$. Let $\hat{r}_y = \text{rank}(\tilde{Y}) = \text{rank}(Y)$. Then with a universal constant C', we have

$$\|\hat{M}\|_{tr} \leq \frac{\sqrt{\rho}\hat{r}_y\sqrt{n_{\max}}}{2C'\mu^2\lambda\check{\kappa}} \quad (2)$$

Proof: Since

$$\hat{M} = (\check{\Lambda}_\mu^\top \check{\Lambda}_\mu)^{-1}\check{\Lambda}_\mu^\top \tilde{Y}$$

Then

$$\begin{aligned}\|\hat{M}\|_{tr} &\leq \|\hat{M}\|_{sp}\hat{r}_y = \|(\check{\Lambda}_\mu^\top \check{\Lambda}_\mu)^{-1}\check{\Lambda}_\mu^\top \tilde{Y}\|_{sp}\hat{r}_y \\ &\leq \|(\check{\Lambda}_\mu^\top \check{\Lambda}_\mu)^{-1}\|_{sp}\|\check{\Lambda}_\mu\|_{sp}\|\tilde{Y}\|_{sp}\hat{r}_y\end{aligned}$$

We use σ_x to denote the largest singular value of $\check{\Lambda}_\mu$, and σ_{xs} to denote the smallest singular value of $\check{\Lambda}_\mu$; thus $\sigma_{xs} \geq \mu\sigma_x$. Based on Lemma 5 of (Chiang, Hsieh, and Dhillon 2015), we then have $\sigma_x \geq C'\lambda\check{\kappa}\sqrt{2N}$. Moreover, based on the definition in (10) and equalities of matrix norms, we have $\|\tilde{Y}\|_{sp} = \|Y\|_{sp} \leq \sqrt{NL}\|Y\|_{max} = \sqrt{\rho NL}$, where the norm $\|Y\|_{max} = \max_{ij}|Y_{ij}|$. Thus

$$\begin{aligned}\|\hat{M}\|_{tr} &\leq \frac{\sigma_x\|\tilde{Y}\|_{sp}\hat{r}_y}{\sigma_{xs}^2} \\ &\leq \frac{\|\tilde{Y}\|_{sp}\hat{r}_y}{\mu^2\sigma_x} \\ &\leq \frac{\|\tilde{Y}\|_{sp}\hat{r}_y}{\mu^2C'\lambda\check{\kappa}\sqrt{2N}} \\ &\leq \frac{\sqrt{\rho NL}\ \hat{r}_y}{\mu^2C'\lambda\check{\kappa}\sqrt{2N}} \\ &\leq \frac{\sqrt{\rho}\hat{r}_y\sqrt{n_{\max}}}{2C'\mu^2\lambda\check{\kappa}}\end{aligned}$$

For low-rank Y, $\hat{r}_y$ is much smaller than the size of Y, this Lemma shows the $\|\hat{M}\|_{tr}$ is upper bounded by $O(\sqrt{n_{\max}})$. Since $\|\hat{M}\|_{tr}$ is obtained from the unconstrained problem without considering the nuclear norm constraint, we can choose $O(\sqrt{n_{\max}})$ as an upper bound for τ.

4. Proof of Proposition 3

Proof: The gradient $\nabla g(M)$ can be computed as

$$\begin{aligned}\nabla g(M) =& 2(\tilde{X}A - \overline{A})^\top H(\tilde{X}A - \overline{A})M \\ &+ 2\rho\overline{A}^\top\left(\Omega \circ (\overline{A}M - Y)\right)\end{aligned}$$

Let $\widehat{\Omega} = [\mathbf{0}_{d,L}; \Omega]$. Then for any two matrices, $M \in \mathbb{R}^{(d+N)\times L}$ and $\widehat{M} \in \mathbb{R}^{(d+N)\times L}$, we have

$$\begin{aligned}&\|\nabla g(M) - \nabla g(\widehat{M})\|_F \\ =& \|2\Gamma(M - \widehat{M}) + 2\rho\left(\widehat{\Omega} \circ (M - \widehat{M})\right)\|_F \\ \leq& 2\|\Gamma(M - \widehat{M})\|_F + 2\rho\|M - \widehat{M}\|_F \\ =& 2\left\|I_L \otimes \Gamma\,\text{Vec}(M - \widehat{M})\right\|_2 + 2\rho\|M - \widehat{M}\|_F \\ \leq& 2\|I_L \otimes \Gamma\|_{sp}\|\,\text{Vec}(M - \widehat{M})\|_2 + 2\rho\|M - \widehat{M}\|_F \\ =& 2\sigma_{\max}(\Gamma)\|M - \widehat{M}\|_F + 2\rho\|M - \widehat{M}\|_F\end{aligned}$$

where $\otimes$ denotes the kronecker product of two matrices; $Vec(\cdot)$ is the vectorization operator; and $\sigma_{\max}(\cdot)$ denotes the largest singular value. It is then straightforward to show η^* is the Lipschitz constant of ∇g with

$$\|\nabla g(M) - \nabla g(\widehat{M})\|_F \leq \eta^*\|M - \widehat{M}\|_F, \text{ for any } M, \widehat{M}.$$

This η^* ensures the update in Algorithm 1 satisfies the conditions of (Beck and Teboulle 2009)[Theorem 4.4], and hence Algorithm 1 has a quadratic convergence rate. ■

References

Beck, A., and Teboulle, M. 2009. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. on Imaging Sciences* 2(1):183–202.

Chiang, K.; Hsieh, C.; and Dhillon, I. 2015. Matrix completion with noisy side information. In *Proc. of NIPS*.