Appendix for "Multi-Omics Data Integration via Supervised Adaptive Sparse Canonical Correlation Analysis"

Heng Zhang, Zhi-Ping Liu

APPENDIX

A. Proof of Theorem 1

In this appendix, we will demonstrate the theoretical convergence of Algorithm 1. The critical point is that the value of the objective function $f\left(A^k,\left\{W_i^k,\delta_i^k\right\}_{i=1}^m\right)$ strictly decreases with iterations. Accordingly, we will show that updates to A, $\{\delta_i\}_{i=1}^m$, and $\{W_i\}_{i=1}^m$ all result in lower objective function value. Before proving Theorem 1, we first give Lemma 1 [1] and Lemma 2 [2].

Lemma 1. *Let $g(x)=x-ax^{\frac{1}{a}}$, where $a\in(0,1)$. Then for any $x>0$, $g(x)\le 1-a$, and $x=1$ is the unique maximizer.*

Lemma 2. *Suppose the compact SVD of matrix $Y=U_YD_YV_Y^\top\in\mathbb{R}^{m\times n}$, where $U_Y\in\mathbb{R}^{m\times r}$, $D_Y\in\mathbb{R}^{r\times r}$, $V_Y\in\mathbb{R}^{n\times r}$, and r is rank of matrix Y. Then the optimal X to the problem*

$$\max_{X\in\mathcal{S}_{m,n}}\quad \operatorname{tr}\left(Y^\top X\right) \tag{A.1}$$

is

$$X=U_YV_Y^\top, \tag{A.2}$$

where $\mathcal{S}_{m,n}$ is Stiefel manifold of $m\times n$.

For the sake of concise expression, let

$$Q_i(A,W_i)=\sum_{k=1,l=1}^{n}\left\|A_{(k,:)}-X_{i(l,:)}W_i\right\|_F^2E_{(k,l)},$$
$$\hat{\delta}_i=f_i(A,W_i)^{-q}.$$

The proof of Theorem 1 is given below.

Proof. According to Lemma 2, we know A^{k+1} is an optimal solution of Eq.(12). Therefore, we can derive

$$\sum_{i=1}^m\delta_i^kQ_i\left(A^{k+1},W_i^k\right)\le\sum_{i=1}^m\delta_i^kQ_i\left(A^k,W_i^k\right), \tag{A.3}$$

which means

$$f\left(A^{k+1},\left\{W_i^k,\delta_i^k\right\}_{i=1}^m\right)\le f\left(A^k,\left\{W_i^k,\delta_i^k\right\}_{i=1}^m\right). \tag{A.4}$$

Based on Algorithm 1, W_i^{k+1} can bring down the objective function. Thus we have

$$\begin{aligned}&Q_i\left(A^{k+1},W_i^{k+1}\right)+\frac{\lambda}{2}\operatorname{tr}\left({W_i^{k+1}}^\top H_i^kW_i^{k+1}\right)\\ \le&Q_i\left(A^{k+1},W_i^k\right)+\frac{\lambda}{2}\operatorname{tr}\left({W_i^k}^\top H_i^kW_i^k\right).\end{aligned} \tag{A.5}$$

For simplifying expression, let $w_j=W_{i(:,j)}$ in this subsection. Similarly to above, by Lemma 1, let $a=1/2$, $x=\left\|w_j^{k+1}\right\|_2/\left\|w_j^k\right\|_2$, the following inequalities hold

$$\frac{\left\|w_j^{k+1}\right\|_2}{\left\|w_j^k\right\|_2}-\frac{1}{2}\frac{\left\|w_j^{k+1}\right\|_2^2}{\left\|w_j^k\right\|_2^2}\le 1-\frac{1}{2}. \tag{A.6}$$

Multiplying Eq.(A.6) by $\left\|w_j^k\right\|_2$ and summing its over j, we have

$$\sum_{j=1}^d\left(\frac{\left\|w_j^{k+1}\right\|_2^2}{\left\|w_j^{k+1}\right\|_2}-\frac{\left\|w_j^{k+1}\right\|_2^2}{2\left\|w_j^k\right\|_2}\right)\le\sum_{j=1}^d\left(\frac{\left\|w_j^k\right\|_2^2}{\left\|w_j^k\right\|_2}-\frac{\left\|w_j^k\right\|_2^2}{2\left\|w_j^k\right\|_2}\right),$$

which means

$$\begin{aligned}&\lambda\operatorname{tr}\left({W_i^{k+1}}^\top H_i^{k+1}W_i^{k+1}\right)-\frac{1}{2}\lambda\operatorname{tr}\left({W_i^{k+1}}^\top H_i^kW_i^{k+1}\right)\\ \le&\lambda\operatorname{tr}\left({W_i^k}^\top H_i^kW_i^k\right)-\frac{1}{2}\lambda\operatorname{tr}\left({W_i^k}^\top H_i^kW_i^k\right).\end{aligned} \tag{A.7}$$

Adding Eq.(A.5) and Eq.(A.7), the following formula will be derived,

$$f_i\left(A^{k+1},W_i^{k+1}\right)\le f_i\left(A^{k+1},W_i^k\right).$$

Accordingly, the following inequality holds

$$f\left(A^{k+1},\left\{W_i^{k+1},\delta_i^k\right\}_{i=1}^m\right)\le f\left(A^{k+1},\left\{W_i^k,\delta_i^k\right\}_{i=1}^m\right). \tag{A.8}$$

Let $f_i^{k+1}:=f_i\left(A^{k+1},W_i^{k+1}\right)$. Noting Lemma 1, let $a=1-q$, $x=\left(f_i^{k+1}/f_i^k\right)^{1-q}$, the following inequalities hold,

$$\frac{\left(f_i^{k+1}\right)^{1-q}}{\left(f_i^k\right)^{1-q}}-\frac{1-q}{2}\frac{f_i^{k+1}}{f_i^k}\le 1-\frac{1-q}{2} \tag{A.9}$$

for $i=1,2,\ldots,m$. Multiplying Eq.(A.9) by $\left(f_i^k\right)^{1-q}$, we have following inequalities simultaneously

$$\frac{f_i^{k+1}}{\left(f_i^{k+1}\right)^q}-\frac{1-q}{2}\frac{f_i^{k+1}}{\left(f_i^k\right)^q}\le\frac{f_i^k}{\left(f_i^k\right)^q}-\frac{1-q}{2}\frac{f_i^k}{\left(f_i^k\right)^q} \tag{A.10}$$

for $i=1,2,\ldots,m$. Further, we can get

$$\begin{aligned}&\sum_{i=1}^m\left(\hat{\delta}_i^{k+1}f_i^{k+1}-\frac{1-q}{2}\hat{\delta}_i^kf_i^{k+1}\right)\\ \le&\sum_{i=1}^m\left(\hat{\delta}_i^kf_i^k-\frac{1-q}{2}\hat{\delta}_i^kf_i^k\right).\end{aligned} \tag{A.11}$$

Combining Eq.(A.4) and Eq.(A.8), we arrive at

$$f\left(A^{k+1},\{W_i^{k+1},\delta_i^k\}_{i=1}^m\right)\le f\left(A^k,\{W_i^k,\delta_i^k\}_{i=1}^m\right) \tag{A.12}$$

and

$$f\left(A^{k+1},\left\{W_i^{k+1},\hat{\delta}_i^k\right\}_{i=1}^m\right)\le f\left(A^k,\left\{W_i^k,\hat{\delta}_i^k\right\}_{i=1}^m\right). \tag{A.13}$$

Combining Eq.(A.11) and Eq.(A.13), we can get

$$\sum_{i=1}^m \hat{\delta}_i^{k+1} f_i^{k+1}\le \sum_{i=1}^m \hat{\delta}_i^k f_i^k. \tag{A.14}$$

Since $\hat{\delta}_i^{k+1}\ge \hat{\delta}_i^k$ for any $i=1,2,\ldots,m$, according to Eq.(A.14) the following inequality holds,

$$f\left(A^{k+1},\{W_i^k,\delta_i^{k+1}\}_{i=1}^m\right)\le f\left(A^k,\{W_i^k,\delta_i^k\}_{i=1}^m\right). \tag{A.15}$$

Based on Eq.(A.15) and Eq.(A.8), we can easily derive the following inequality

$$f\left(A^{k+1},\{W_i^{k+1},\delta_i^{k+1}\}_{i=1}^m\right)\le f\left(A^k,\{W_i^k,\delta_i^k\}_{i=1}^m\right).$$

Consequently, the sequence $\left\{f\left(A^k,\{W_i^k,\delta_i^k\}_{i=1}^m\right)\right\}$ generated by Algorithm 1 converges since $f\left(A,\{W_i,\delta_i\}_{i=1}^m\right)$ has a lower bound. □

B. Proof of Theorem 2

To prove Theorem 2, we first introduce the following Lemma.

Lemma 3. *For any $0<\epsilon<1$, with probability $1-\epsilon$, the following inequality holds,*

$$\left|\mathcal{L}-\hat{\mathcal{L}}\right|\le \frac{8Mm^2}{\sqrt{n}}+2M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}, \tag{A.16}$$

where $\mathcal{L}$ and $\hat{\mathcal{L}}$ are defined in Eq.(25) and Eq.(26), respectively.

Proof. Inspired by [3], for any samples $S=\{x_1,x_2,\ldots,x_n\}$, let S' be samples different from S by only one instance x_r'. The empirical risk of the hypothesis function on S' is denoted as $\hat{\mathcal{L}}'$. We have

$$\begin{aligned}
&\left|\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}\right|-\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}'\right|\right|\\
\le&\sup_{h\in\mathcal{H}}\left|\hat{\mathcal{L}}-\hat{\mathcal{L}}'\right|\\
\le&\sup_{h\in\mathcal{H}}\frac{1}{n^2}\left(\begin{array}{l}\sum_{i=1}^m\left(\left|h_i\left(x_r,x_r\right)\right|+\left|h_i\left(x_r',x_r'\right)\right|\right)\\ +\sum_{\substack{i=1\\j=1}}^m\sum_{l=1}^n\left(\left|h_{i,j}\left(x_l,x_r\right)\right|+h_{i,j}\left|\left(x_l,x_r'\right)\right|\right)\end{array}\right)\\
\le&\frac{2M}{n^2}\left(m+m^2n\right)\\
\le&\frac{2M\left(m^2+1\right)}{n},
\end{aligned}$$

where the last inequality is due to the fact that h is bounded by M and $m<n$. According to the McDiarmid inequality [4], we have the following inequality holds with probability $1-\epsilon$,

$$\left|\mathcal{L}-\hat{\mathcal{L}}\right|\le +\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}\right|+2M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.17}$$

Then we analyze the upper bound of the expectation term, i.e. $\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}\right|$. First, we have

$$\begin{aligned}
&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}\right|\\
=&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\begin{array}{l}\mathcal{L}-\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n h_i\left(x_l,x_l\right)\\ -\frac{1}{n^2}\sum_{\substack{i=1\\j=1}}^m\sum_{\substack{l=1\\s=1}}^n h_{i,j}\left(x_l,x_s\right)\end{array}\right|\\
\le&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}_1-\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n h_i\left(x_l,x_l\right)\right|\\
&+\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}_2-\frac{1}{n^2}\sum_{\substack{i=1\\j=1}}^m\sum_{\substack{l=1\\s=1}}^n h_{i,j}\left(x_l,x_s\right)\right|,
\end{aligned}$$

where $\mathcal{L}_1=\sum_{i=1}^m\mathbb{E}_x h_i\left(x,x\right)$ and $\mathcal{L}_2=\sum_{i=1,j=1}^m\mathbb{E}_{x,y}h_{i,j}\left(x,y\right)$. Let $\sigma_1,\sigma_2,\ldots,\sigma_n$ be i.i.d. Rademacher random variables taking values in $\{-1,1\}$ with equal probability and $\bar{S}=\{\bar{x}_1,\bar{x}_2,\ldots,\bar{x}_n\}$ be an independent copy of S. Then the first term can be bounded by

$$\begin{aligned}
&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}_1-\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n h_i\left(x_l,x_l\right)\right|\\
\le&\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n\left(h_i\left(x_l,x_l\right)-h_i\left(\bar{x}_l,\bar{x}_l\right)\right)\right|\\
=&\mathbb{E}_{S,\bar{S},\sigma}\sup_{h\in\mathcal{H}}\left|\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n\sigma_l\left(h_i\left(x_l,x_l\right)-h_i\left(\bar{x}_l,\bar{x}_l\right)\right)\right|\\
=&2\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\frac{1}{n^2}\sum_{i=1}^m\sum_{l=1}^n\sigma_l h_i\left(x_l,x_l\right)\right|\\
\le&\frac{2m}{n^2}\max_i\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^n\sigma_l h_i\left(x_l,x_l\right)\right|.
\end{aligned}$$

The second term can be bounded by

$$\begin{aligned}
&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}_2-\frac{1}{n^2}\sum_{\substack{i=1\\j=1}}^m\sum_{\substack{l=1\\s=1}}^n h_{i,j}\left(x_l,x_s\right)\right|\\
=&\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{n^2}\sum_{\substack{i=1\\j=1}}^m\sum_{\substack{l=1\\s=1}}^n\left(h_{i,j}\left(x_l,x_s\right)-h_{i,j}\left(\bar{x}_l,\bar{x}_s\right)\right)\right|\\
\le&\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{n\left(n-1\right)}\sum_{\substack{i=1\\j=1}}^m\sum_{l\neq s}^n\left(h_{i,j}\left(x_l,x_s\right)-h_{i,j}\left(\bar{x}_l,\bar{x}_s\right)\right)\right|
\end{aligned}$$

$$
\begin{aligned}
&+\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{\substack{i=1\\j=1}}^{m}\sum_{l=1}^{n}\left(h_{i,j}\left(x_l,x_l\right)-h_{i,j}\left(\bar{x}_l,\bar{x}_l\right)\right)\right|\\
\leq&\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{\substack{i=1\\j=1}}^{m}\sum_{l=1}^{\lfloor n/2\rfloor}\begin{pmatrix}h_{i,j}\left(x_l,x_{l+\lfloor n/2\rfloor}\right)\\-h_{i,j}\left(\bar{x}_l,\bar{x}_{l+\lfloor n/2\rfloor}\right)\end{pmatrix}\right|\\
&+\mathbb{E}_{S,\bar{S}}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{\substack{i=1\\j=1}}^{m}\sum_{l=1}^{n}\left(h_{i,j}\left(x_l,x_l\right)-h_{i,j}\left(\bar{x}_l,\bar{x}_l\right)\right)\right|\\
=&2\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{\substack{i=1\\j=1}}^{m}\sum_{l=1}^{\lfloor n/2\rfloor}\sigma h_{i,j}\left(x_l,x_{l+\lfloor n/2\rfloor}\right)\right|\\
&+2\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{\substack{i=1\\j=1}}^{m}\sum_{l=1}^{n}\sigma h_{i,j}\left(x_l,x_l\right)\right|\\
\leq&\frac{2m^2}{\lfloor n/2\rfloor}\max_{i,j}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^{\lfloor n/2\rfloor}\sigma h_{i,j}\left(x_l,x_{l+\lfloor n/2\rfloor}\right)\right|\\
&+\frac{2m^2}{n}\max_{i,j}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^{n}\sigma h_{i,j}\left(x_l,x_l\right)\right|,
\end{aligned}
$$

where the second inequality is obtained by the Lemma A.1 in [5]. Combining the above results, according to the Khintchine-Kahane inequality [6], we have

$$
\begin{aligned}
&\mathbb{E}_S\sup_{h\in\mathcal{H}}\left|\mathcal{L}-\hat{\mathcal{L}}\right|\\
\leq&\frac{2m}{n^2}\max_{i}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^{n}\sigma_l h_i\left(x_l,x_l\right)\right|\\
&+\frac{2m^2}{\lfloor n/2\rfloor}\max_{i,j}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^{\lfloor n/2\rfloor}\sigma h_{i,j}\left(x_l,x_{l+\lfloor n/2\rfloor}\right)\right|\\
&+\frac{2m^2}{n}\max_{i,j}\mathbb{E}_{S,\sigma}\sup_{h\in\mathcal{H}}\left|\sum_{l=1}^{n}\sigma h_{i,j}\left(x_l,x_l\right)\right|\\
\leq&\frac{2m}{n^2}\max_{i}\mathbb{E}_{S}\sup_{h\in\mathcal{H}}\left(\sum_{l=1}^{n}h_i\left(x_l,x_l\right)^2\right)^{\frac{1}{2}}\\
&+\frac{2m^2}{\lfloor n/2\rfloor}\max_{i,j}\mathbb{E}_{S}\sup_{h\in\mathcal{H}}\left(\sum_{l=1}^{\lfloor n/2\rfloor}h_{i,j}\left(x_l,x_{l+\lfloor n/2\rfloor}\right)^2\right)^{\frac{1}{2}}\\
&+\frac{2m^2}{n}\max_{i,j}\mathbb{E}_{S}\sup_{h\in\mathcal{H}}\left(\sum_{l=1}^{n}h_{i,j}\left(x_l,x_l\right)^2\right)^{\frac{1}{2}}\\
\leq&\frac{2Mm}{n\sqrt{n}}+\frac{2Mm^2}{\sqrt{\lfloor n/2\rfloor}}+\frac{2Mm^2}{\sqrt{n}}\leq\frac{8Mm^2}{\sqrt{n}}.
\end{aligned}
$$

Incorporating this bound into Eq.(A.17), we have the following inequality holds with probability $1-\epsilon$,

$$
\left|\mathcal{L}-\hat{\mathcal{L}}\right|\leq\frac{8Mm^2}{\sqrt{n}}+2M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.18}
$$

□

The proof of Theorem 2 is given below.

Proof. Let $\hat{\mathcal{L}}^*$ and $\hat{\mathcal{L}}_f^*$ are optimal empirical risk for the models corresponding adaptive and fixed weights, and the expectation risk are $\mathcal{L}^*$ and $\mathcal{L}_f^*$, respectively. According to Lemma 3, with probability $1-\epsilon$, the following inequalities holds,

$$
\mathcal{L}^*-\hat{\mathcal{L}}^*\leq\frac{8Mm^2}{\sqrt{n}}+2M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}, \tag{A.19a}
$$

$$
\hat{\mathcal{L}}_f^*-\mathcal{L}_f^*\leq\frac{8Mm^2}{\sqrt{n}}+2M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.19b}
$$

Combing Eq.(A.19a) and Eq.(A.19b), the following inequality holds with probability $1-\epsilon$,

$$
\mathcal{L}^*+\hat{\mathcal{L}}_f^*-\hat{\mathcal{L}}^*\leq\mathcal{L}_f^*+\frac{16Mm^2}{\sqrt{n}}+4M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.20}
$$

Notice that since the the δ_i in $\mathcal{L}_f^*$ is fixed, yet the δ_i in $\mathcal{L}^*$ is required to be optimized, as thus $\hat{\mathcal{L}}_f^*-\hat{\mathcal{L}}^*\geq 0$. Accordingly, with at least probability $1-\epsilon$, we have

$$
\mathcal{L}^*\leq\mathcal{L}_f^*+\frac{16Mm^2}{\sqrt{n}}+4M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.21}
$$

□

C. *The proof of Corollary 1*

Notably, Since the definition Eq.(10) of the weights shows that δ_i cannot be taken to 1 or 0 for all i, Corollary 1 needs to be proved in the sense of a limit. This part of the proof relies heavily on the inequality between two sequences preserved in the limit, i.e.,

Lemma 4. *Let both $\{a_n\}$ and $\{b_n\}$ be convergent sequences. If there exists a positive integer N such that $a_n<b_n$ when $n>N$, then $\lim_{n\to\infty}a_n\leq\lim_{n\to\infty}b_n$.*

The proof of Corollary 1 is given below.

Proof. Without loss of generality, let the weights be $\left\{\delta_i=1-\sum_{j\neq i}\delta_j,\delta_j=1/n|j\neq i\right\}$, then the $\left\{\hat{\mathcal{L}}_{i(n)}^*\right\}$ can be viewed a sequence about n. According to subsection B, we have $\hat{\mathcal{L}}_{i(n)}^*\geq\hat{\mathcal{L}}^*$ for any $n\in\mathbb{N}$. Clearly, $\hat{\mathcal{L}}_{i(n)}^*$ converges to $\hat{\mathcal{L}}_i^*$ when n tends to infinity. Thus by Lemma 4, it follows that $\hat{\mathcal{L}}_i^*\geq\hat{\mathcal{L}}^*$ holds. Furthermore, similar to Eq.(A.21), we can get the following inequality holds with at least probability $1-\epsilon$,

$$
\mathcal{L}^*\leq\mathcal{L}_i^*+\frac{16Mm^2}{\sqrt{n}}+4M\left(m^2+1\right)\sqrt{\frac{\log\frac{2}{\epsilon}}{2n}}. \tag{A.22}
$$

□

References

[1] L. Wang and S. Chen, "Joint representation classification for collective face recognition," *Pattern Recognition*, vol. 63, pp. 182–192, 2017.

[2] W. Zhuge, T. Luo, H. Tao, C. Hou, and D. Yi, "Multi-view spectral clustering with incomplete graphs," *IEEE Access*, vol. 8, pp. 99 820–99 831, 2020.

[3] S. Li and Y. Liu, "Sharper generalization bounds for clustering," in *International Conference on Machine Learning*. PMLR, pp. 6392–6402.

[4] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[5] S. Clémençon, G. Lugosi, and N. Vayatis, "Ranking and empirical minimization of U-statistics," *The Annals of Statistics*, vol. 36, no. 2, pp. 844–874, 2008.

[6] R. Latała and K. Oleszkiewicz, "On the best constant in the Khinchin-Kahane inequality," *Studia Mathematica*, vol. 109, no. 1, pp. 101–104, 1994.