1. Supplementary material

Lemma 1 (Lemma 14 in Appendix) *Let G_1 and G_2 be two zero mean Gaussian distributions with covariance matrix $\mathbf{\Gamma\Sigma\Gamma}$ and $\mathbf{\Gamma\Theta\Gamma}$. Furthermore $\mathbf{\Sigma}$ and $\mathbf{\Theta}$ are positive definite matrices. If there exists (i,j) such that*

$$|\mathbf{\Sigma}_{i,j} - \mathbf{\Theta}_{i,j}| \geq \delta(\mathbf{\Sigma}_{i,i} + \mathbf{\Theta}_{i,i} + \mathbf{\Sigma}_{j,j} + \mathbf{\Theta}_{j,j}), \tag{1}$$

then the total variation distance between G_1 and G_2 is at least $\frac{1}{12e^{1/4}}\delta$.

Proof Given $\phi_1(u)$ and $\phi_2(u)$ as characteristic function of G_1 and G_2 respectively. Due to Lemma 2 in (Moridomi et al., 2018), we have

$$\int_x |G_1(x) - G_2(x)|dx \geq \max_{u\in\mathbb{R}^N} |\phi_1(u) - \phi_2(u)|, \tag{2}$$

so we only need to show the lower bound of $\max_{u\in\mathbb{R}^N} |\phi_1(u) - \phi_2(u)|$.

Then we set that characteristic function of G_1 and G_2 are $\phi_1(u) = e^{\frac{-1}{2}u^T\mathbf{\Gamma}^T\mathbf{\Sigma\Gamma}u}$ and $\phi_2(u) = e^{\frac{-1}{2}u^T\mathbf{\Gamma}^T\mathbf{\Theta\Gamma}u}$ respectively. Set that $\alpha_1 = (\mathbf{\Gamma}v)^T\mathbf{\Sigma}(\mathbf{\Gamma}v)$, $\alpha_2 = (\mathbf{\Gamma}v)^T\mathbf{\Theta}(\mathbf{\Gamma}v)$ and $\mathbf{\Gamma}u = \frac{\mathbf{\Gamma}v}{\sqrt{\alpha_1+\alpha_2}}$. Moreover we denote that $\bar{v} = \Gamma v$, for any $\bar{v} \in \mathbb{R}^V$, there exists $v \in \mathbb{R}^V$. $\bar{u} = \Gamma u$ in the same way.

We need only give the lower bound of $\max_{u\in\mathbb{R}^N} |\phi_1(u) - \phi_2(u)|$.

Next we have that

$$\begin{aligned}
&\max_{u\in\mathbb{R}^N} |\phi_1(u) - \phi_2(u)| \\
&= \max_{u\in\mathbb{R}^N} \left| e^{\frac{-1}{2}u^T\mathbf{\Gamma\Sigma\Gamma}u} - e^{\frac{-1}{2}u^T\mathbf{\Gamma\Theta\Gamma}u} \right| \\
&= \max_{u\in\mathbb{R}^V} \left| e^{\frac{-1}{2}(\mathbf{\Gamma}u)^T\mathbf{\Sigma}(\mathbf{\Gamma}u)} - e^{\frac{-1}{2}(\mathbf{\Gamma}u)^T\mathbf{\Theta}(\mathbf{\Gamma}u)} \right| \\
&\geq \max_{\bar{v}\in\mathbb{R}^N} \left| e^{\frac{-\alpha_1}{2(\alpha_1+\alpha_2)}} - e^{\frac{-\alpha_2}{2(\alpha_1+\alpha_2)}} \right| \\
&\geq \max_{\bar{v}\in\mathbb{R}^N} \left| \frac{1}{2e^{1/4}} \frac{\alpha_1 - \alpha_2}{\alpha_1 + \alpha_2} \right|.
\end{aligned} \tag{3}$$

Then second inequality is due to Lemma 5, since $\min\{\frac{\alpha_1}{\alpha_1+\alpha_2}, \frac{\alpha_2}{\alpha_1+\alpha_2}\} \in (0, \frac{1}{2}]$.

Due to assumption in the Lemma we obtain for some (i,j) that

$$\begin{aligned}
&\delta(\mathbf{\Sigma}_{i,i} + \mathbf{\Theta}_{i,i} + \Sigma_{j,j} + \mathbf{\Theta}_{j,j}) \leq |\mathbf{\Sigma}_{i,j} - \mathbf{\Theta}_{i,j}| \\
&= \frac{1}{2}|(\boldsymbol{e}_i + \boldsymbol{e}_j)^T(\mathbf{\Sigma} - \mathbf{\Theta})(\boldsymbol{e}_i + \boldsymbol{e}_j) - \boldsymbol{e}_i^T(\mathbf{\Sigma} - \mathbf{\Theta})\boldsymbol{e}_i - \boldsymbol{e}_j^T(\mathbf{\Sigma} - \mathbf{\Theta})\boldsymbol{e}_j|
\end{aligned} \tag{4}$$

It implies that one of $(\boldsymbol{e}_i+\boldsymbol{e}_j)^T(\mathbf{\Sigma}-\mathbf{\Theta})(\boldsymbol{e}_i+\boldsymbol{e}_j)$, $\boldsymbol{e}_i^T(\mathbf{\Sigma}-\mathbf{\Theta})\boldsymbol{e}_i$ and $\boldsymbol{e}_j^T(\mathbf{\Sigma}-\mathbf{\Theta})\boldsymbol{e}_j$ has absolute value greater that $\frac{2\delta}{3}(\mathbf{\Sigma}_{i,i} + \mathbf{\Theta}_{i,i} + \mathbf{\Sigma}_{j,j} + \mathbf{\Theta}_{j,j})$.

Since Σ, Θ are strictly positive definite matrices, we have that for all $v \in \{\boldsymbol{e}_i+\boldsymbol{e}_j, \boldsymbol{e}_i, \boldsymbol{e}_j\}$

$$v^T(\mathbf{\Sigma} + \mathbf{\Theta})v \leq 2(\mathbf{\Sigma} + \mathbf{\Theta})_{i,i} + 2(\mathbf{\Sigma} + \mathbf{\Theta})_{j,j}. \tag{5}$$

and therefore we have that

$$\max_{\bar{v}\in\mathbb{R}^N}\left|\frac{1}{2e^{1/4}}\frac{\alpha_1-\alpha_2}{\alpha_1+\alpha_2}\right| \geq \max_{\bar{v}\in\{\boldsymbol{e}_i+\boldsymbol{e}_j,\boldsymbol{e}_i,\boldsymbol{e}_j\}}\left|\frac{1}{2e^{1/4}}\frac{v^T(\boldsymbol{\Sigma}-\boldsymbol{\Theta})v}{v^T(\boldsymbol{\Sigma}+\boldsymbol{\Theta})v}\right| \geq \frac{\delta}{6e^{1/4}} \tag{6}$$

■

Now we give the proof of the Main Theorem as follows:

Proof [Proof of Theorem 2] Due to Lemma 4 (in main part) we obtain that

$$\text{Regret}_{\text{OSDP}}(T,\mathcal{K},\mathcal{L},\boldsymbol{W}^*) \leq \frac{H_0}{\eta}+\frac{\eta}{s}T. \tag{7}$$

Due to the main proposition in main part we know that $s = 1/(1152(\beta+\rho\epsilon)^2\sqrt{e}g^2)$.

Thus we need only to show $H_0 \leq \frac{\tau}{\epsilon}$. Given $\boldsymbol{W}_0$ and $\boldsymbol{W}_1$ is the minimizer and maximizer of R respectively, then we obtain that

$$\begin{aligned}
&\max_{\boldsymbol{W},\boldsymbol{W}'\in\mathcal{K}}(R(\boldsymbol{W})-R(\boldsymbol{W}')) = R(\boldsymbol{W}_1)-R(\boldsymbol{W}_0)\\
&= -\ln\det(\boldsymbol{\Gamma}\boldsymbol{W}_1\boldsymbol{\Gamma}+\epsilon\boldsymbol{E})+\ln\det(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma}+\epsilon\boldsymbol{E})\\
&= \sum_{i=1}^N \ln\frac{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma})+\epsilon}{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_1\boldsymbol{\Gamma})+\epsilon}\\
&= \sum_{i=1}^N \ln\left(\frac{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma})}{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_1\boldsymbol{\Gamma})+\epsilon}+\frac{\epsilon}{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_1\boldsymbol{\Gamma})+\epsilon}\right)\\
&\leq \sum_{i=1}^N \ln\left(\frac{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma})}{\epsilon}+1\right)\\
&\leq \sum_{i=1}^N \frac{\lambda_i(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma})}{\epsilon} = \frac{\text{Tr}(\boldsymbol{\Gamma}\boldsymbol{W}_0\boldsymbol{\Gamma})}{\epsilon} \leq \frac{\tau}{\epsilon}.
\end{aligned} \tag{8}$$

Plugging s, we obtain that

$$\text{Regret}_{\text{OSDP}}(T,\mathcal{K},\mathcal{L},\boldsymbol{W}^*) = O\left(g^2(\beta+\rho\epsilon)^2T\eta+\frac{\tau}{\epsilon\eta}\right). \tag{9}$$

■

Lemma 2 (Lemma A.1 (Moridomi et al., 2018)) *Let P and Q be probability distributions over $\mathbb{R}^N$ and $\phi_P(u)$ and $\phi_Q(u)$ be their characteristic functions, respectively. Then*

$$\max_{u\in\mathbb{R}^N}|\phi_P(u)-\phi_Q(u)| \leq \int_x |P(x)-Q(x)|dx, \tag{10}$$

the right hand side is the total variation distance between any distribution Q and P.

Lemma 3 (Lemma A.2 (Christiano, 2014)) *Let P and Q be probability distributions over $\mathbb{R}^N$ with total variation distance δ. Then*

$$H(\alpha P + (1-\alpha)Q) \leq \alpha H(P) + (1-\alpha)H(Q) - \alpha(1-\alpha)\delta^2, \tag{11}$$

where $H(P) = \mathbb{E}_{x\sim P}[\ln P(x)]$.

Lemma 4 (Lemma A.3 (Moridomi et al., 2018)) *For any probability distribution P over $\mathbb{R}^N$ with zero mean and covariance matrix $\mathbf{\Sigma}$, its entropy is bounded by the log-determinant of covariance matrix. That is*

$$-H(P) \leq \frac{1}{2}\ln(\det(\Sigma)(2\pi e)^N). \tag{12}$$

Lemma 5 (Lemma A.4 (Moridomi et al., 2018))

$$e^{\frac{-x}{2}} - e^{-\frac{1-x}{2}} \geq \frac{e^{-1/4}}{2}(1-2x), \tag{13}$$

for $0 \leq x \leq 1/2$.

2. Definition of biclustered structure and ideal quasi dimension

As in Herbster et al. (2020), we define the class of (k,l)-biclustered structure matrices as follows:

Definition 6 *For $m \geq k$ and $n \geq l$, the class of (k,l)-binary biclustered matrices is defined as*

$$\mathbb{B}_{k,l}^{m\times n} = \{\boldsymbol{U} \in \{-1,+1\}^{m\times n} : \boldsymbol{r} \in [k]^m, \boldsymbol{c} \in [l]^n, \boldsymbol{V} \in \{1,-1\}^{k\times l}, \boldsymbol{U}_{i,j} = \boldsymbol{V}_{r_i,c_j}, i \in [m], j \in [n]\}.$$

Denote $\mathcal{B}^{m,d} = \{\boldsymbol{R} \subset \{0,1\}^{m\times d} : \|\boldsymbol{R}_i\|_2 = 1, i \in [m], \mathrm{rank}(\boldsymbol{R}) = d\}$, for any matrix $\boldsymbol{U} \in \mathbb{B}_{k,l}^{m,n}$ we can decompose $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{U}^*\boldsymbol{C}^\top$ for some $\boldsymbol{U}^* \in \{-1,+1\}^{k\times l}, \boldsymbol{R} \in \mathcal{B}^{m,k}$ and $\boldsymbol{C} \in \mathcal{B}^{n,l}$.

Theorem 7 ((Herbster et al., 2020)) *If $\boldsymbol{U} \in \mathbb{B}_{k,l}^{m\times n}$ define $\mathcal{D}^o_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$ as*

$$\mathcal{D}^o_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) = 2\mathrm{Tr}(\boldsymbol{R}^\top \boldsymbol{M}\boldsymbol{R})\alpha_{\boldsymbol{M}} + 2\mathrm{Tr}(\boldsymbol{C}^\top \boldsymbol{N}\boldsymbol{C})\alpha_{\boldsymbol{N}} + 2k + 2l, \tag{14}$$

where $\boldsymbol{M}, \boldsymbol{N}$ are PD-Laplacian, as the minimum over all decompositions of $\boldsymbol{U} = \boldsymbol{R}\boldsymbol{U}^*\boldsymbol{C}^\top$ for some $\boldsymbol{U}^* \in \{-1,+1\}^{k\times l}, \boldsymbol{R} \in \mathcal{B}^{m,k}$ and $\boldsymbol{C} \in \mathcal{B}^{n,l}$. Thus, for $\boldsymbol{U} \in \mathbb{B}_{k,l}^{m\times n}$,

$$\mathcal{D}^\gamma_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) \leq \mathcal{D}^o_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}), \tag{15}$$

if $\|\boldsymbol{U}\|_{\max} \leq \frac{1}{\gamma}$.

Moreover, we define the max-norm of a matrix $\boldsymbol{U} \in \mathbb{R}^{m\times n}$ as follows:

$$\|\boldsymbol{U}\|_{\max} = \min_{\boldsymbol{P}\boldsymbol{Q}^\top=\boldsymbol{U}} \left\{ \max_{1\le i\le m} \|\boldsymbol{P}_i\| \max_{1\le j\le n} \|\boldsymbol{Q}_j\| \right\}. \tag{16}$$

Furthermore we define the quasi-dimension of a matrix $\boldsymbol{U}$ with $\boldsymbol{M} \in \mathbb{S}^{m\times m}_{++}$ and $\boldsymbol{N} \in \mathbb{S}^{n\times n}_{++}$ at margin γ as

$$\mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) = \min_{\bar{\boldsymbol{P}}\bar{\boldsymbol{Q}}^\top=\gamma\boldsymbol{U}} \alpha_{\boldsymbol{M}} \mathrm{Tr}(\bar{\boldsymbol{P}}^\top \boldsymbol{M}\bar{\boldsymbol{Q}}) + \alpha_{\boldsymbol{N}} \mathrm{Tr}(\bar{\boldsymbol{Q}}^\top \boldsymbol{N}\bar{\boldsymbol{Q}}). \tag{17}$$

See section 4.1 from Herbster et al. (2020), if $\boldsymbol{U}$ is a (k,l)-biclustered structured matrix, they show an example where $\mathcal{D}^{o}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U}) \in O(k+l)$ with ideal side information. When exactly that there exists a sequence that $y_t = (\bar{\boldsymbol{P}}\bar{\boldsymbol{Q}}^\top)_{i_t,j_t} = \boldsymbol{U}_{i_t,j_t}$ where $(\bar{\boldsymbol{P}},\bar{\boldsymbol{Q}}) = \arg\min_{\boldsymbol{P},\boldsymbol{Q}} \mathcal{D}^{\gamma}_{\boldsymbol{M},\boldsymbol{N}}(\boldsymbol{U})$, and $\boldsymbol{U}$ satisfies the assumptions in Herbster et al. (2020), then we have that $\widehat{\mathcal{D}} \in O(k+l)$ with same side information.

3. Online similarity prediction with side information

In this section, we show the application of our reduction method and generalised log-determinant regularizer to online similarity prediction with side information.

Let $G = (V, E)$ be an undirected and connected graph with $n = |V|$ vertices and $m = |E|$ edges. Assign vertices to K classes with a vector $\boldsymbol{y} = \{y_1, \cdots, y_n\}$ where $y_i \in \{1, \cdots, K\}$. For a matrix $\boldsymbol{L}$, we denote $\boldsymbol{L}^+$ as pseudo-inverse matrix of $\boldsymbol{L}$. The online similarity prediction is defined as follows: On each round t, for a given pair of vertices (i_t, j_t) algorithm needs to predict whether they are in the same class denoted as $\hat{y}_{i_t,j_t}$. If they are in the same class then $y_{i_t,j_t} = 1$, $y_{i_t,j_t} = -1$, otherwise. Our target is to give a bound of the prediction mistakes $M = \sum_{t=1}^{T} \mathbb{I}_{\hat{y}_{i_t,j_t} \neq y_{i_t,j_t}}$.

Definition 8 *The set of cut-edges in $(G, \boldsymbol{y})$ is denoted as $\Phi^G(\boldsymbol{y}) = \{(i,j) \in E : y_i \neq y_j\}$ we abbreviate it to Φ^G and the cut-size is given as $|\Phi^G(\boldsymbol{y})|$. The set of cut-edges with respect to class label k is denoted as $\Phi^G_k(\boldsymbol{y}) = \{(i,j) \in E : k \in \{y_i, y_j\}, y_i \neq y_j\}$. Note that $\sum_{s=1}^{k} |\Phi^G_s(\boldsymbol{y})| = 2|\Phi^G(\boldsymbol{y})|$. Given $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ such that $\boldsymbol{A}_{ij} = \boldsymbol{A}_{ji} = 1$ if $(i,j) \in E(G)$ and $\boldsymbol{A}_{ij} = 0$, otherwise. $\boldsymbol{D}$ is denoted as diagonal matrix with $\boldsymbol{D}_{ii}$ is the degree of vertex i. We define the Laplacian as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A}$.*

Definition 9 *If G is identified with a resistive network such that each edge is a unit resistor, then the effective resistance $R^G_{i,j}$ between pair $(i,j) \in V^2$ can be defined as $R^G_{i,j} = (e_i - e_j)\boldsymbol{L}^+(e_i - e_j)$, where e_i is the i-th vector in the canonical basis of $\mathbb{R}^n$.*

Gentile et al. (2013) gave a mistake bound in the following proposition:

Proposition 10 *Let $(G, \boldsymbol{y})$ be a labeled graph. If we run the Matrix Winnow with G as input graph, we have the following mistake bound*

$$M^W = O\left(|\Phi^G| \max_{(i,j)\in V^2} R^G_{i,j} \ln n \right) \tag{18}$$

In our new reduction, we define the comparator matrix $\boldsymbol{U} \in \{1,-1\}^{n\times n}$ where if vertices i,j are in the same class then $\boldsymbol{U}_{ij}=1$, and $\boldsymbol{U}_{ij}=-1$, otherwise. Firstly, we denote that $\mathbf{1}$ is a K-dimensional vector that all entries are 1. Due to (Gentile et al., 2013; Herbster et al., 2020), we see that $\boldsymbol{U}$ is a (K,K)-biclustered $n\times n$ matrix where $\boldsymbol{U}^* = 2\boldsymbol{I}_K - \mathbf{1}\mathbf{1}^\top$, and there exists $\boldsymbol{R}\in\mathcal{B}^{n,k}$ such that $\boldsymbol{U}=\boldsymbol{R}\boldsymbol{U}^*\boldsymbol{R}^\top$. Define the side information matrices $\boldsymbol{M}=\boldsymbol{N}\in\mathbb{R}^{n\times n}$ as PD-Laplcian $\tilde{\boldsymbol{L}}$, where $\boldsymbol{L}$ is the Laplacian matrix based on the graph G.

Thus we have

$$\boldsymbol{\Gamma} = \begin{bmatrix} \sqrt{\alpha_{\widetilde{\boldsymbol{L}}}\widetilde{\boldsymbol{L}}} & 0 \\ 0 & \sqrt{\alpha_{\widetilde{\boldsymbol{L}}}\widetilde{\boldsymbol{L}}} \end{bmatrix}, \tag{19}$$

where $\alpha_{\tilde{\boldsymbol{L}}} = \max_i(\tilde{\boldsymbol{L}}^{-1})_{ii}$.

According to Herbster et al. (2020), we further obtain that $\frac{1}{\gamma}\in O(1)$, more concisely we can set that $\frac{1}{\gamma}=3$. Meanwhile given sparse matrix $\boldsymbol{Z}$ in the following equation

$$\boldsymbol{Z}\langle i,j\rangle = \frac{1}{2}(\boldsymbol{e}_i\boldsymbol{e}_{n+j}^\top + \boldsymbol{e}_{n+j}\boldsymbol{e}_i^\top). \tag{20}$$

Thus we give the following proposition for our reduction from a graph based online similarity prediction to a generalised OSDP problem $(\mathcal{K},\mathcal{L})$ with bounded $\boldsymbol{\Gamma}$-trace norm.

Proposition 11 *Given an online similarity prediction problem with graph $(G,\boldsymbol{y})$, then we can reduce this problem to a generalised OSDP problem $(\mathcal{K},\mathcal{L})$ with bounded $\boldsymbol{\Gamma}$-trace norm such that*

$$\mathcal{K} = \left\{\boldsymbol{X}\in\mathbb{S}^{n\times n}_{++} : |\boldsymbol{X}_{ii}|\le 1, \mathrm{Tr}(\boldsymbol{\Gamma}\boldsymbol{X}\boldsymbol{\Gamma})\le\widehat{\mathcal{D}}\right\}$$
$$\mathcal{L} = \{c\boldsymbol{Z}\langle i,j\rangle : c\in\{-1/\gamma, 1/\gamma\}, i\in[n], j\in[n]\},$$

where $\boldsymbol{\Gamma}$ is defined as above, and $\widehat{\mathcal{D}}$ is arbitrary. In particular, we have that

$$M = \sum_{t=1}^{T} \mathbb{I}_{\hat{y}_{i_t,j_t}\neq y_{i_t,j_t}} \le \mathrm{Regret}_{\mathrm{OSDP}}(M,\mathcal{K},\mathcal{L}) \tag{21}$$

According to Herbster et al. (2020), there exists $\bar{\boldsymbol{P}},\bar{\boldsymbol{Q}}$ such that $\boldsymbol{U}^*=\bar{\boldsymbol{P}}\bar{\boldsymbol{Q}}^\top$, it implies that the hinge loss $\mathrm{hloss}(\mathcal{S},\gamma)=0$.

Remark 12 *According to Theorem 3 and section 4.2 in (Herbster et al., 2020) if $\boldsymbol{U}$ obtains the (K,K)-biclustered structure, i.e., there exists $\boldsymbol{U}^*$, such that $\boldsymbol{U}^*=2\boldsymbol{I}_K-\mathbf{1}\mathbf{1}^\top$, and there exists $\boldsymbol{R}\in\mathcal{B}^{n,k}$ such that $\boldsymbol{U}=\boldsymbol{R}\boldsymbol{U}^*\boldsymbol{R}^\top$, due to our Theorem 13 in main part, we have that*

$$M \le O\left(\mathrm{Tr}(\boldsymbol{R}^\top\boldsymbol{L}\boldsymbol{R})\alpha_{\boldsymbol{L}}\right), \tag{22}$$

where $\boldsymbol{L}$ is Laplacian of the corresponding graph G.

Remark 13 *According to Herbster et al. (2020), we have that*

$$\mathrm{Tr}(\boldsymbol{R}^\top \boldsymbol{L}\boldsymbol{R}) \le 2 \sum_{(i,j)\in E} \|\boldsymbol{R}_i - \boldsymbol{R}_j\|^2,$$

where $\sum_{(i,j)\in E} \|\boldsymbol{R}_i - \boldsymbol{R}_j\|^2$ *counts only when there is an edge between different classes. Due to the definition of* $|\Phi^G|$*, we have that* $\sum_{(i,j)\in E} \|R_i - R_j\|^2 = |\Phi^G|$*.*

Simultaneously, $\alpha_{\boldsymbol{L}} = \max_{i\in[n]} \boldsymbol{L}^+_{ii}$ *so we obtain that* $\alpha_{\boldsymbol{L}} \ge \boldsymbol{e}_i^\top \boldsymbol{L}^+ \boldsymbol{e}_i, \forall i \in [n]$*. It implies that* $4\alpha_{\boldsymbol{L}} \ge \max_{(i,j)\in V^2} R^G_{i,j}$*. Thus we have that our new mistake bound improves the previous bound a logarithmic factor and recovers the previous bound up to a constant factor.*

References

Paul Christiano. Online local learning via semidefinite programming. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 468–474. ACM, 2014.

Claudio Gentile, Mark Herbster, and Stephen Pasteris. Online similarity prediction of networked data from known and unknown graphs. In *Conference on Learning Theory*, pages 662–695, 2013.

Mark Herbster, Stephen Pasteris, and Lisa Tse. Online matrix completion with side information. *Advances in Neural Information Processing Systems*, 33, 2020.

Ken-ichiro Moridomi, Kohei Hatano, and Eiji Takimoto. Online linear optimization with the log-determinant regularizer. *IEICE Transactions on Information and Systems*, 101 (6):1511–1520, 2018.