Perceptronic Complexity and Online Matrix Completion

Stephen Pasteris SPASTERIS@TURING.AC.UK
The Alan Turing Institute, London, UK

Editors: Shipra Agrawal and Francesco Orabona

Abstract

We present an online algorithm for learning a binary relation, or equivalently the components of a binary-valued matrix. We derive mistake bounds for this algorithm that are based on a novel complexity measure called the perceptronic complexity. Informally, if we consider each row of the matrix as a separate learning task, then the perceptronic complexity is the Novikoff perceptron bound for learning the whole matrix when given the optimal kernel over the columns of the matrix. Our mistake bound is equal, up to a logarithmic factor, to the perceptronic complexity of the matrix plus the perceptronic complexity of its transpose. We show how this mistake bound asymptotically outperforms those of previous algorithms for the same problem, and give examples of our bound on natural families of matrices.

Keywords: Online Learning, Matrix Completion, Multitask Learning

1. Introduction and Related Work

Binary matrix completion is a learning task with numerous real world applications. In this task we have an unknown binary relationship $\sim$ between two sets enumerated as $[m]$ and $[n]$. The aim is to predict, when given $(i,j) \in [m] \times [n]$, whether $i \sim j$ or not. In this paper we consider the online protocol in which learning proceeds in trials. On each trial t we are given a pair $(i_t, j_t) \in [m] \times [n]$ and have to predict whether $i_t \sim j_t$ or not, after which we learn whether we were right or wrong. The aim is to minimise the number of mistakes. For a detailed overview of the online protocol and online algorithms see Cesa-Bianchi and Lugosi (2006) and Shalev-Shwartz (2012).

In this paper we represent the binary relationship $\sim$ by an $m \times n$ dimensional matrix $\boldsymbol{U}$ where $u_{i,j} := 1$ if $i \sim j$ and $u_{i,j} := 0$ otherwise. We define a new notion of complexity of a binary matrix, which we call the 'perceptronic complexity'. Informally, given that each row of a binary matrix is considered as a different learning task, then the perceptronic complexity of that matrix is the Novikoff perceptron bound (Rosenblatt (1958), Novikoff (1963)) for learning the entire matrix when given the optimal kernel over the columns. We present an algorithm for the above learning problem in which the number of mistakes is bounded above, up to a factor logarithmic in $m+n$, by the perceptronic complexity of $\boldsymbol{U}$ or that of its transpose.

The general problem of matrix completion has been studied extensively in the batch statistical i.i.d. setting, see for example Srebro and Shraibman (2005), Candès and Tao (2010), Pontil and Maurer (2013) and references therein. We note that standard online-to-batch conversion techniques allow the online algorithm presented in this paper to be applied to the batch setting also. Algorithms for the online setting were first given in Goldman et al. (1989) and Goldman and Warmuth (1993) but worked with a limited relation/matrix model. Less limited algorithms were then given in Hazan et al. (2012) and Herbster et al. (2016). We show that our bound improves over the bounds of these algorithms, in doing so implying that we inherit bounds (via Foygel (2012)) based on the max norm and trace norm defined in Srebro and Shraibman (2005), which are commonly used measures of

© 2023 S. Pasteris.

matrix complexity used in matrix completion algorithms. Our algorithm utilises the algorithm of Herbster et al. (2020), which was itself based on Warmuth (2007), as a subroutine. It should be noted that Moridomi et al. (2018) managed to remove (by modifying the algorithm) a logarithmic factor appearing in the bound of Hazan et al. (2012) - it is an open problem as to whether the same can be done to our work.

We give two examples of relationship/matrix model along with derived bounds specific to them. These are a latent factor model, which generalises the biclustering model of Herbster et al. (2016), and a similarity model, which was essentially given in Herbster et al. (2016). Our bounds for these models significantly improve on those given in Herbster et al. (2016). The analysis of our similarity model is based on work in Herbster et al. (2008) and Herbster and Pontil (2006). We give real world examples of the models: medical diagnosis, recommender systems, and online dating.

Our paper is structured as follows. In Section 2 we give the definitions of the notation that we use throughout the paper. In Section 3 we define the perceptronic complexity of a matrix and give an equivalent formulation. In Section 4 we define the learning problem and give our mistake bound. In Section 5 we compare our result to those of Herbster et al. (2016) and Hazan et al. (2012). In Section 6 we give our example matrix models along with bounds on their perceptronic complexities. In Section 7 we define and describe our algorithm. Finally, in Appendix A we prove all the theorems in the paper, in order of appearance.

2. Definitions

Given $k \in \mathbb{N}$ define $[k] := \{1, 2, \cdots, k\}$. Given a predicate p we define $[\![p]\!] := 1$ if p is true and $[\![p]\!] := 0$ otherwise. Scalars and sets will be denoted by non-bold lower and upper case letters respectively. Vectors and matrices will be denoted by bold lower and upper case latin letters respectively. Given a matrix denoted by some bold upper case letter, its i-th row vector will be denoted by the equivalent bold lower case letter with subscript i. e.g. if a matrix is denoted by $\boldsymbol{X}$, its i-th row vector is denoted by $\boldsymbol{x}_i$ and its (i,j)-th component is denoted $x_{i,j}$. Given a vector $\boldsymbol{x}$ we define $\|\boldsymbol{x}\|$ to be its L_2 norm. Given a matrix $\boldsymbol{X}$ we denote its transpose by $\boldsymbol{X}^\top$ and let x_* be the minimiser of $|x_{i,j}|$ across all rows i and columns j. Given some $l \in \mathbb{N}$ we let $\mathcal{P}^l$ be the set of all positive semidefinite $l \times l$ dimensional matrices. Given $l, k \in \mathbb{N}$ and a matrix $\boldsymbol{X} \in \mathbb{R}^{l \times k}$ (which could be a vector ($k = 1$) or scalar ($l, k = 1$)) we define $\text{sign}(\boldsymbol{X})$ to be the matrix $\boldsymbol{X'} \in \mathbb{R}^{l \times k}$ such that $x'_{i,j} = [\![x_{i,j} \geq 0]\!] - [\![x_{i,j} < 0]\!]$ for all $(i,j) \in [l] \times [k]$. Given matrices $\boldsymbol{X} \in \mathbb{R}^{p \times l}$ and $\boldsymbol{X'} \in \mathbb{R}^{q \times l}$ (which are vectors if $l = 1$) for some $p, q, l \in \mathbb{N}$ we define $[\boldsymbol{X}, \boldsymbol{X'}]$ to be equal to the matrix $\boldsymbol{X''} \in \mathbb{R}^{(p+q) \times l}$ such that $\boldsymbol{x}''_i = \boldsymbol{x}_i$ and $\boldsymbol{x}''_{p+j} = \boldsymbol{x}'_j$ for all $(i,j) \in [p] \times [q]$. Given $k, l \in \mathbb{N}$ let $\boldsymbol{0}^{k,l}$ be the $k \times l$-dimensional matrix in which every component is equal to 0. Given $k, l \in \mathbb{N}$ let $\mathcal{R}^{k,l}$ be the set of all $\boldsymbol{X} \in \mathbb{R}^{k \times l}$ with $\|\boldsymbol{x}_i\| = 1$ for all $i \in [k]$. Given a set S let 2^S be the set of all subsets of S.

3. Perceptronic Complexity

In this section we define and analyse our measure of complexity of a binary matrix: the perceptronic complexity. Viewing each row of a matrix as a seperate learning task, the perceptronic complexity of that matrix is the Novikoff bound of the perceptron algorithm for learning the matrix components, given the optimal kernel over the columns. We now formally define this quantity. To do this we first introduce some common notions in machine learning.

Given some $l \in \mathbb{N}$ and a positive semidefinite matrix $\boldsymbol{K} \in \mathcal{P}^l$ (a.k.a. 'kernel matrix') we define the radius of $\boldsymbol{K}$ as:

$$\rho(\boldsymbol{K}) := \max_{i \in [l]} \sqrt{k_{i,i}}\,.$$

Given, in addition, a vector $\boldsymbol{u} \in \{-1, 1\}^l$ we define the set of (normalised) linear separators of $(\boldsymbol{K}, \boldsymbol{u})$ as:

$$\Lambda(\boldsymbol{K}, \boldsymbol{u}) := \{\boldsymbol{w} \in \mathbb{R}^l \mid \operatorname{sign}(\boldsymbol{K}\boldsymbol{w}) = \boldsymbol{u}\ ;\ \boldsymbol{w}^\top \boldsymbol{K}\boldsymbol{w} = 1\}$$

and define the margin of $(\boldsymbol{K}, \boldsymbol{u})$ as

$$\delta(\boldsymbol{K}, \boldsymbol{u}) := \max_{\boldsymbol{w} \in \Lambda(\boldsymbol{K}, \boldsymbol{u})} \min_{i \in [n]} |\boldsymbol{w} \cdot \boldsymbol{k}_i|$$

where the maximum of the empty set is 0. We note that $\rho(\boldsymbol{K})^2/\delta(\boldsymbol{K}, \boldsymbol{u})^2$ is the Novikoff bound of the perceptron algorithm when learning the vector $\boldsymbol{u}$ given kernel $\boldsymbol{K}$.

As their names imply, the radius and margin have a geometric meaning. Every kernel $\boldsymbol{K} \in \mathcal{P}^l$ can be represented (non-uniquely) by a set of l points in a real vector space, where $k_{i,j}$ is the inner product between points i and j. Inversely, every set of l points in a real vector space can be represented by an unique kernel. Given such a set of points (and hence a kernel $\boldsymbol{K}$), the radius of the kernel is the minimum radius of a ball, centred at the origin, that contains all the points. Given a vector $\boldsymbol{u} \in \{-1, 1\}^l$ which labels each point i with u_i, the margin of $(\boldsymbol{K}, \boldsymbol{u})$ is defined as half the maximum distance between two parallel hyperplanes, equidistant from the origin, such that no points lie between the hyperplanes, and the set of points lying on the far side of a hyperplane are either the set of points labelled 1 or the set of points labelled -1.

Given a matrix $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$ for some $m, n \in \mathbb{N}$ we define its perceptronic complexity as:

$$\tau(\boldsymbol{U}) := \min_{\boldsymbol{K} \in \mathcal{P}^n} \sum_{i \in [m]} \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2}$$

which is the sum of Novikoff bounds of each row of $\boldsymbol{U}$ when given the optimal kernel over the columns of $\boldsymbol{U}$.

We described above how a kernel can be represented by a set of points in a real vector space. This representation leads to the following way of quantifying the perceptronic complexity. First define, for all $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$, the set:

$$\mathcal{N}(\boldsymbol{U}) := \{(\boldsymbol{P}, \boldsymbol{Q}) \mid \exists\, l \in \mathbb{N} : (\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{R}^{m,l} \times \mathcal{R}^{n,l}\ ;\ \forall (i,j) \in [m] \times [n],\ \operatorname{sign}(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) = u_{i,j}\}$$

which is the set of all possible row-normalised factorisations of matrices that have the same sign pattern as $\boldsymbol{U}$. We then have the following equality:

Theorem 1 *Given $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$ we have:*

$$\tau(\boldsymbol{U}) = \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})} \sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 .$$

4. Online Matrix Completion

In this paper we solve the following learning problem. We have an unknown matrix $\boldsymbol{U} \in \{-1,1\}^{m\times n}$. Learning proceeds in trials $t = 1, 2, 3, \cdots, T$. On trial t:

1. Some $(i_t, j_t) \in [m] \times [n]$ is revealed.
2. We select $\hat{y}_t \in \{-1, 1\}$.
3. $y_t := u_{i_t,j_t}$ is revealed.

The aim is to minimise the total number of mistakes:

$$\mathcal{M} := \sum_{t\in[T]} [\![\hat{y}_t \neq y_t]\!] \,.$$

We will give, in Section 7, an algorithm PCMC (Perceptronic Complexity Matrix Completion), with running time $\mathcal{O}(m^3 + n^3)$ per trial, whose mistakes are bounded as follows:

Theorem 2 *The number of mistakes $\mathcal{M}$ incurred by* PCMC *is bounded above by:*

$$\mathcal{M} \in \mathcal{O}\left((\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))\ln(m+n)\right) .$$

5. Analysing the Improvement

For our problem Herbster et al. (2016) gave a mistake bound of:

$$\mathcal{M} \in \mathcal{O}\left(\ln(m+n)(m+n) \min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{N}(\boldsymbol{U})} \max_{(i,j)\in[m]\times[n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2\right)$$

Since an average is no greater than a maximum, Theorem 1 shows that our bound in Theorem 2 is never outperformed by this bound and is often significantly better.

We now compare our mistake bound to that of Hazan et al. (2012) on the same problem. In order to do this we make the following definitions. For any matrix $\boldsymbol{U} \in \{-1,1\}^{m\times n}$ we define its sign-consistent set as:

$$\mathcal{S}(\boldsymbol{U}) := \{\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times n} \mid \text{sign}(\tilde{\boldsymbol{U}}) = \boldsymbol{U}\} \,.$$

For any matrix $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times n}$, we define its symmetrisation $\text{sym}(\tilde{\boldsymbol{U}}) \in \mathbb{R}^{(m+n)\times(m+n)}$ as:

$$\text{sym}(\tilde{\boldsymbol{U}}) := \left[\left[\boldsymbol{0}^{m,m}, \tilde{\boldsymbol{U}}^\top\right]^\top, \left[\tilde{\boldsymbol{U}}, \boldsymbol{0}^{n,n}\right]^\top\right]$$

and its semi-definite decomposition set as:

$$\mathcal{D}(\tilde{\boldsymbol{U}}) := \{(\boldsymbol{A}, \boldsymbol{B}) \mid \boldsymbol{A}, \boldsymbol{B} \in \mathcal{P}^{(m+n)} \,;\, \boldsymbol{A} - \boldsymbol{B} = \text{sym}(\tilde{\boldsymbol{U}})\} \,.$$

Given $\boldsymbol{A}, \boldsymbol{B} \in \mathcal{P}^{(m+n)}$ we define:

$$\tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) := \max\left\{\max_{i\in[m+n]} a_{i,i} \,,\, \max_{i\in[m+n]} b_{i,i}\right\} \quad ; \quad \tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) := \sum_{i\in[m+n]} a_{i,i} + \sum_{i\in[m+n]} b_{i,i} \,.$$

The algorithm of Hazan et al. (2012) solves a more general problem than ours. Specifically, instead of choosing $\hat{y}_t \in \{-1, 1\}$ on each trial t we now choose some $w_t \in [-1, 1]$ and then receive a convex g-Lipschitz loss function $\ell_t : [-1, 1] \to \mathbb{R}$ instead of $y_t \in \{-1, 1\}$. Given that we only choose to update on a subset of trials $Z \subseteq [T]$, the bound given in Hazan et al. (2012) is then, for any $\tilde{\boldsymbol{U}} \in [-1, 1]^{m \times n}$, given by:

$$\sum_{t \in Z} \ell_t(w_t) \leq \sum_{t \in Z} \ell_t(\tilde{u}_{i_t, j_t}) + 2g \min_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})} \sqrt{\tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) \tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) \ln(2(m+n)) |Z|}\,. \tag{1}$$

This bound leads to the following theorem.

Theorem 3 *Given $m, n \in \mathbb{N}$, $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$, $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$ and $\{(i_t, j_t) \mid t \in [T]\}$ let $g := 1/\tilde{u}_*$ and for all $t \in [T]$ let $y_t := u_{i_t, j_t}$ and $\ell_t(x) := \max\{0, -g y_t x + 1\}$. Then given Z and $\{w_t \mid t \in [T]\} \subseteq [-1, 1]$ are such that $Z = \{t \in [T] \mid y_t \neq \operatorname{sign}(w_t)\}$ and Equation (1) is satisfied we have:*

$$\sum_{t \in [T]} [\![y_t \neq \operatorname{sign}(w_t)]\!] \in \mathcal{O}\left(\frac{\ln(m+n)}{\tilde{u}_*^2} \min_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})} \tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) \tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) \right).$$

Hence, the algorithm of Hazan et al. (2012) leads to an algorithm for our problem whose mistakes are bounded by the right hand side of the inequality in Theorem 3. We now show that our bound is a two-fold improvement over this. To do this we first define, for a matrix $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m \times n}$, its factorisation set as:

$$\mathcal{F}(\tilde{\boldsymbol{U}}) := \{(\boldsymbol{P}, \boldsymbol{Q}) \mid \exists\, l \in \mathbb{N} :\ \boldsymbol{P} \in \mathbb{R}^{m \times l}\ ;\ \boldsymbol{Q} \in \mathbb{R}^{n \times l}\ ;\ \boldsymbol{P}\boldsymbol{Q}^\top = \tilde{\boldsymbol{U}}\}\,.$$

We now rewrite the bound of Theorem 3 in terms of a factorisation.

Theorem 4 *Given a matrix $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m \times n}$ we have:*

$$\min_{(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})} \tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) \tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) \in \Theta\left(\min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 \max_{j \in [n]} \|\boldsymbol{q}_j\|^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 \right) \right)$$

Theorem 4 implies that the minimiser, over all $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$, of the bound in Theorem 3 is in $\Theta(\ln(m+n) h(\boldsymbol{U}))$, where:

$$h(\boldsymbol{U}) := \min_{\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})} \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 \max_{j \in [n]} \frac{\|\boldsymbol{q}_j\|^2}{\tilde{u}_*^2} + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \max_{i \in [m]} \frac{\|\boldsymbol{p}_i\|^2}{\tilde{u}_*^2} \right).$$

So to compare our bound to that of Hazan et al. (2012) we must compare $\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top)$ to $h(\boldsymbol{U})$. To do this we first define $h'(\boldsymbol{U})$ by replacing each appearance of $\tilde{u}_*$ in $h(\boldsymbol{U})$ with $\tilde{u}_{i,j}$, noting that this can only decrease the quantity. i.e:

$$h'(\boldsymbol{U}) := \min_{\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})} \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 \max_{j \in [n]} \left(\frac{\|\boldsymbol{q}_j\|}{\tilde{u}_{i,j}} \right)^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \max_{i \in [m]} \left(\frac{\|\boldsymbol{p}_i\|}{\tilde{u}_{i,j}} \right)^2 \right).$$

The next theorem shows that $h'(\boldsymbol{U})$ can be a polynomial factor less than $h(\boldsymbol{U})$:

Theorem 5 *For all $k \in \mathbb{N}$ there exists a matrix $\boldsymbol{U} \in \{-1,1\}^{(k+k^2)\times(k+k^2)}$ with:*

$$h(\boldsymbol{U}) \in \Omega(k^{5/2}) \quad \text{and} \quad h'(\boldsymbol{U}) \in \mathcal{O}(k^2)\,.$$

So there exists a family of matrices $\boldsymbol{U}$ with $h'(\boldsymbol{U}) \in \mathcal{O}(h(\boldsymbol{U})^{4/5})$. We do not know if one can improve on Theorem 5. i.e. it may be the case in which families of matrices $\boldsymbol{U}$ exist such that $h'(\boldsymbol{U}) \in o(h(\boldsymbol{U})^{4/5})$. As far as we are aware, being able to bound $\mathcal{M}$ by $\mathcal{O}(h'(\boldsymbol{U})\ln(m+n))$ is itself a novel result. However, we further improve on this. To see this further improvement we first rewrite $h'(\boldsymbol{U})$ via the following theorem:

Theorem 6 *For all $\boldsymbol{U} \in \{-1,1\}^{m\times n}$ we have:*

$$h'(\boldsymbol{U}) = \min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{N}(\boldsymbol{U})} \left(\sum_{i\in[m]} \max_{j\in[n]} \left(\frac{1}{\boldsymbol{p}_i\cdot\boldsymbol{q}_j}\right)^2 + \sum_{j\in[n]} \max_{i\in[m]} \left(\frac{1}{\boldsymbol{p}_i\cdot\boldsymbol{q}_j}\right)^2 \right).$$

By Theorem 1 we have:

$$\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top) = \min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{N}(\boldsymbol{U})} \sum_{i\in[m]} \max_{j\in[n]} \left(\frac{1}{\boldsymbol{p}_i\cdot\boldsymbol{q}_j}\right)^2 + \min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{N}(\boldsymbol{U})} \sum_{j\in[n]} \max_{i\in[m]} \left(\frac{1}{\boldsymbol{p}_i\cdot\boldsymbol{q}_j}\right)^2.$$

Note the difference between $h'(\boldsymbol{U})$, as written in Theorem 6, and $(\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))$ written as above - in $h'(\boldsymbol{U})$ we have a single pair $(\boldsymbol{P},\boldsymbol{Q})$ whilst in $(\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))$ we use separate pairs for the two summations, which can only decrease the quantity. It is an open problem to quantify just how much smaller $(\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))$ can be than $h'(\boldsymbol{U})$.

6. Examples

In this section we give bounds, for example families of matrices $\boldsymbol{U} \in \mathbb{R}^{m\times n}$, of the main term of our mistake bound: $\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top)$. To aid us in doing this we make the following definition. For all matrices $\boldsymbol{U} \in \{-1,1\}^{m\times n}$ let:

$$\mathcal{E}(\boldsymbol{U}) := \left\{ \boldsymbol{C} \in \mathbb{R}^{m\times n} \;\middle|\; \exists(\boldsymbol{P},\boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U}) : \forall (i,j) \in [m]\times[n]\,,\; c_{i,j} \geq \left(\frac{1}{\boldsymbol{p}_i\cdot\boldsymbol{q}_j}\right)^2 \right\}.$$

By Theorem 1 we have:

$$\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top) = \min_{\boldsymbol{C}\in\mathcal{E}(\boldsymbol{U})} \sum_{i\in[m]} \max_{j\in[n]} c_{i,j} + \min_{\boldsymbol{D}\in\mathcal{E}(\boldsymbol{U})} \sum_{j\in[n]} \max_{i\in[m]} d_{i,j} \tag{2}$$

so in this section we will give results about the existence of matrices $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ which have bounds on $c_{i,j}$ for all $(i,j) \in [m]\times[n]$. To start, the next two theorems show that, given one or two matrices in $\mathcal{E}(\boldsymbol{U})$, there exists a matrix in $\mathcal{E}(\boldsymbol{U})$ which can lead to better asymptotic bounds.

Theorem 7 *Given $\boldsymbol{D},\boldsymbol{F} \in \mathcal{E}(\boldsymbol{U})$ there exists $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ such that for all $(i,j) \in [m]\times[n]$ we have:*

$$c_{i,j} \in \mathcal{O}(\min\{d_{i,j}, f_{i,j}\})\,.$$

Theorem 8 *Given $\boldsymbol{F} \in \mathcal{E}(\boldsymbol{U})$ there exists $\boldsymbol{C}, \boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ such that for all $i \in [m]$ we have:*

$$\max_{j\in[n]} c_{i,j} \in \mathcal{O}\left(\min_{z\in\{-1,1\}} \max_{j\in[n]} [\![u_{i,j} = z]\!] f_{i,j}\right)$$

and for all $j \in [n]$ we have:

$$\max_{i\in[m]} d_{i,j} \in \mathcal{O}\left(\min_{z\in\{-1,1\}} \max_{i\in[m]} [\![u_{i,j} = z]\!] f_{i,j}\right).$$

Now that we have these theorems in hand we give existence results about matrices in $\mathcal{E}(\boldsymbol{U})$ for two different models of matrix $\boldsymbol{U}$. Combining these existence results with theorems 7 and 8 and Equation (2) gives us a wide range of mistake bounds, significantly improving over those in Herbster et al. (2016).

6.1. The Latent Factor Model

Our first example is for a latent factor model. In this model we have k latent factors. For each latent factor l there is a set $A_l \subseteq [m]$ of objects 'with' the latent factor and a set $B_l \subseteq [n]$ of 'symptoms' of the latent factor. We then have $u_{i,j} = 1$ if and only if object i displays symptom j. Although noise can easily be incorporated into our bounds we here only consider, for simplicity, the noise-free case.

In addition to the obvious medical application (where the set of objects $[m]$ are patients and the latent factors are underlying illnesses) this model also has applications in recommender systems, where each latent factor l is a genre of item, A_l is the set of users that like that genre, and B_l is the set of items in that genre.

Mistake bounds for the latent factor model can be derived from the following theorem and Equation (2), possibly utilising theorems 7 and 8.

Theorem 9 *Suppose we have some $k, m, n \in \mathbb{N}$ and collections of sets:*

$$\{A_l \mid l \in [k]\} \subseteq 2^{[m]} \text{ and } \{B_l \mid l \in [k]\} \subseteq 2^{[n]}.$$

Define the matrix $\boldsymbol{U} \in \mathbb{R}^{m\times n}$ by:

$$\forall (i,j) \in [m] \times [n],\ u_{i,j} := 2[\![\exists\, l \in [k] : (i,j) \in A_l \times B_l]\!] - 1.$$

Then there exists $\boldsymbol{C}, \boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ with:

$$c_{i,j} \in \mathcal{O}\left(\sum_{l\in[k]} [\![i \in A_l]\!] \max_{j'\in B_l} \sum_{l'\in[k]} [\![j' \in B_{l'}]\!]\right) \quad \text{and} \quad d_{i,j} \in \mathcal{O}\left(\sum_{l\in[k]} [\![j \in B_l]\!] \max_{i'\in A_l} \sum_{l'\in[k]} [\![i' \in A_{l'}]\!]\right)$$

for all $(i,j) \in [m] \times [n]$.

An elegant corollary of this is that, given each object $i \in [m]$ has at most a latent factors and each symptom $j \in [n]$ is a symptom of at most b latent factors, we have that the number of mistakes of PCMC is bounded by:

$$\mathcal{M} \in \tilde{\mathcal{O}}\left((m+n) + b\sum_{l\in[k]} |A_l| + a\sum_{l\in[k]} |B_l|\right)$$

A special case of the latent factor model is that in which the (indices of the) rows of $\boldsymbol{U}$ are partitioned into k clusters such that for each cluster all rows in that cluster are identical. By defining $\{A_l \mid l \in [k]\}$ to be this partition we can then define $\{B_l \mid l \in [k]\}$ to be such that B_l is the set of column indices j such that $u_{i,j} = 1$ for all $i \in A_l$. Note that for any $l \in [k]$ we have:

$$\max_{i' \in A_l} \sum_{l' \in [k]} [\![i' \in A_{l'}]\!] = 1$$

and hence, by Theorem 9, there exists $\boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ with:

$$d_{i,j} \in \mathcal{O}\left(\sum_{l \in [k]} [\![j \in B_l]\!] \right) .$$

Substituting this into Equation 2 gives a mistake bound of:

$$\mathcal{M} \in \tilde{\mathcal{O}}\left((m+n) + m \max_{j \in [n]} \sum_{l \in [k]} [\![j \in B_l]\!] + \sum_{l \in [k]} |B_l| \right)$$

When the matrix $\boldsymbol{U}$ is sparse this bound can be a significant improvement over the bound of $\tilde{\mathcal{O}}(k(m+n))$ given by Herbster et al. (2016).

6.2. The Similarity Model

This model is best explained by an application - that of online dating. Here we have two (possibly overlapping) groups of people, the first group is of cardinality m and the second group of cardinality n. Our matrix $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$ is defined so that for all $(i, j) \in [m] \times [n]$ we have $u_{i,j} = 1$ if a date between the i-th person in the first group and the j-th person in the second group would be successful. Some people are similar to each other in terms of personality and it's natural to assume that if person i gets on with person j then he/she is likely to get on with people similar to person j as well. Our bounds for this model are given in terms of a tree over the first group and/or a tree over the second group. This/these trees are those that best capture the similarity between people, in that if two people are linked in a tree then they are likely to be similar. The above assumption then translates to the assumption that if two people are linked in a tree then they are likely to get on with many of the same people.

Although Herbster et al. (2016) considered all graphs rather than just trees, we note that we could also use general graphs instead of just trees. However the bounds for general graphs scale with the resistance diameter and are never better than our results for trees by more than a poly-logarithmic factor. The proofs of our results for trees work by constructing a larger tree from the original tree. This could also be done in Herbster et al. (2016) but our bounds significantly improve on that.

Mistake bounds for the similarity model can be derived from the following theorem and Equation (2), possibly utilising theorems 7 and 8.

Theorem 10 *Suppose we have $m, n \in \mathbb{N}$ and a matrix $\boldsymbol{U} \in \{-1, 1\}^{m \times n}$. Then given trees R and S with vertex sets $[n]$ and $[m]$ respectively, there exists $\boldsymbol{C}, \boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ with:*

$$c_{i,j} \in \mathcal{O}\left(\ln(n)^2 \sum_{(l,k) \in R} [\![u_{i,l} \neq u_{i,k}]\!] \right) \quad \text{and} \quad d_{i,j} \in \mathcal{O}\left(\ln(m)^2 \sum_{(l,k) \in S} [\![u_{l,j} \neq u_{k,j}]\!] \right)$$

for all $(i,j) \in [m] \times [n]$. *Note that the summations are over the edges of the trees.*

We note that the technology in Herbster et al. (2012) can be used to (sometimes) substantially reduce the terms $\ln(n)^2$ and $\ln(m)^2$ that appear in Theorem 10. However, how to do this is beyond the scope of the paper and left as an exercise to the reader.

A particularly elegant corollary of Theorem 10 is that when $\boldsymbol{U}$ is symmetric we have, for any tree R with vertex set $[m]$, that the number of mistakes made by PCMC is bounded by:

$$\mathcal{M} \in \tilde{\mathcal{O}}\left((m+n) + \sum_{(l,k)\in R} \sum_{i\in[m]} [\![u_{l,i} \neq u_{k,i}]\!]\right)$$

where the term $\sum_{i\in[m]} [\![u_{l,i} \neq u_{k,i}]\!]$ is the Hamming distance between rows/columns l and k. Given rows/columns l and k with $(l,k) \in R$, we have that l and k are hopefully similar and hence their Hamming distance is hopefully small - leading to a low mistake bound.

7. Algorithm

In this section we introduce our algorithm PCMC. For all $i \in [m+n]$ we let $\boldsymbol{e}_i$ be the i-th $(m+n)$-dimensional basis vector. In the pseudocode the $:=$ operation denotes the creation of a constant whilst the $\leftarrow$ operation denotes the (re)setting of a variable. We have flags that take values in $\{\circ, \bullet\}$. Our pseudocode is as follows:

PCMC

1. Initialise by:
   (a) Define $d := \lceil \log_2(m+n) \rceil$; $f := 2$; $g := 4$; $c := 3 - e$
   (b) For all $\ell \in [d]$ define $\theta_\ell := \sqrt{1/f^\ell}$; $\pi_\ell := 2gf^\ell \ln(m+n)/c$
   (c) For all $(\ell, \dagger) \in [d] \times \{\circ, \bullet\}$ set $\boldsymbol{W}_\ell^\dagger \leftarrow \boldsymbol{I}$
   (d) For all $i \in [m]$ set $\lambda_i^\circ \leftarrow 1$; $\mu_i^\circ \leftarrow 0$
   (e) For all $j \in [n]$ set $\lambda_j^\bullet \leftarrow 1$; $\mu_j^\bullet \leftarrow 0$
2. For $t = 1, 2, 3 \cdots T$
   (a) Receive (i_t, j_t)
   (b) If $\lambda_{i_t}^\circ \leq \lambda_{j_t}^\bullet$ then $\kappa_t := i_t$; $\ell_t := \lambda_{i_t}^\circ$; $\dagger_t := \circ$. Else $\kappa_t := j_t$; $\ell_t := \lambda_{j_t}^\bullet$; $\dagger_t := \bullet$
   (c) Set $\boldsymbol{X}_t := \frac{1}{2}(\boldsymbol{e}_{i_t} + \boldsymbol{e}_{m+j_t})(\boldsymbol{e}_{i_t} + \boldsymbol{e}_{m+j_t})^\top$
   (d) Output $\hat{y}_t := \text{sign}(\text{Tr}(\boldsymbol{X}_t \boldsymbol{W}_{\ell_t}^{\dagger_t}) - 1)$
   (e) Receive y_t
   (f) If $\hat{y}_t \neq y_t$ then:
      i. Set $\boldsymbol{W}_{\ell_t}^{\dagger_t} \leftarrow \exp(\ln(\boldsymbol{W}_{\ell_t}^{\dagger_t}) + y_t \theta_{\ell_t} \boldsymbol{X}_t)$
      ii. If $\mu_{\kappa_t}^{\dagger_t} < \pi_{\ell_t}$ then $\mu_{\kappa_t}^{\dagger_t} \leftarrow \mu_{\kappa_t}^{\dagger_t} + 1$. Else $\mu_{\kappa_t}^{\dagger_t} \leftarrow 0$; $\lambda_{\kappa_t}^{\dagger_t} \leftarrow \lambda_{\kappa_t}^{\dagger_t} + 1$

We now describe the algorithm, which is inspired by the classic 'doubling trick' (see e.g. Cesa-Bianchi and Lugosi (2006)) for automatically tuning parameters. Define $d := \lceil \log_2(m+n) \rceil$, $f := 2$, $g := 4$ and $c := 3 - e$.

On any trial t we have a flag $\dagger_t \in \{\circ, \bullet\}$. We will describe how $\dagger_t$ is selected later. At any point in time, each $i \in [m]$ has a 'row-level' λ_i° and each $j \in [n]$ has a 'column-level' $\lambda_j^\bullet$. The row-level of i is increased by one when there have been $\pi_\ell := 2gf^\ell \ln(m+n)/c$ mistakes on trials t with $i_t = i$, $\dagger_t = \circ$ and i at the current row-level. Similarly, the column-level of j is increased by one when there have been π_ℓ mistakes on trials t with $j_t = j$, $\dagger_t = \bullet$ and j at the current column-level. Note that the row-levels of elements of $[m]$ and the column-levels of elements of $[n]$ are non-decreasing.

We now describe how $\dagger_t$ and another value $\ell_t \in [d]$ are chosen on trial t. First, the row-level of i_t is compared to the column level of j_t. If the former is no larger than the later then $\dagger_t := \circ$ and ℓ_t is set equal to the row-level of i_t. Otherwise $\dagger_t := \bullet$ and ℓ_t is set equal to the column-level of j_t.

Note that we can partition $[T]$ as:

$$[T] = \bigcup \{L_\ell^\dagger \mid \ell \in [d]\,,\, \dagger \in \{\circ, \bullet\}\}$$

where we define:

$$L_\ell^\dagger := \{t \in [T] \mid \ell_t = \ell\,;\, \dagger_t = \dagger\}$$

for all $\ell \in [d]$ and $\dagger \in \{\circ, \bullet\}$. We now describe how the algorithm behaves on trials in each of the sets in this partition. So suppose we have some $\ell \in [d]$ and $\dagger \in \{\circ, \bullet\}$. Let $S := |L_\ell^\dagger|$. For all $s \in [S]$ let $t(s)$ be the s-th trial in $L_\ell^\dagger$ - i.e. such that $t(s) \in L_\ell^\dagger$ and $\sum_{t' \in L_\ell^\dagger} [\![t' \leq t(s)]\!] = s$. For all $s \in [S]$ let $\tilde{\boldsymbol{W}}_s$ be the matrix $\boldsymbol{W}_\ell^\dagger$ at the start of trial $t(s)$ and let $\tilde{\boldsymbol{X}}_s := \boldsymbol{X}_{t(s)}$. Then the values of $\{\hat{y}_t : t \in L_\ell^\dagger\}$ that are produced by PCMC are equal to those formed by the following algorithm:

1. $\tilde{\boldsymbol{W}}_1 \leftarrow \boldsymbol{I}$
2. For $s = 1, 2, 3 \cdots S$:
   (a) Set $\tilde{\boldsymbol{X}}_s := \frac{1}{2}(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})^\top$
   (b) Output $\hat{y}_{t(s)} := \mathrm{sign}(\mathrm{Tr}(\tilde{\boldsymbol{X}}_s \tilde{\boldsymbol{W}}_s) - 1)$
   (c) $\tilde{\boldsymbol{W}}_{s+1} := \exp(\ln(\tilde{\boldsymbol{W}}_s) + \frac{1}{2}\theta_\ell(y_{t(s)} - \hat{y}_{t(s)})\tilde{\boldsymbol{X}}_s)$

This algorithm is essentially that of Herbster et al. (2016) with parameter θ_ℓ. We will refer to this algorithm as MCMC (Margin Complexity Matrix Completion). PCMC hence works with $2d$ instances of MCMC in parallel: one for each pair in $[d] \times \{\circ, \bullet\}$. The current state of the instance of MCMC corresponding to the pair $(\ell, \dagger)$ is stored in the matrix $\boldsymbol{W}_\ell^\dagger$. In order to bound the mistakes of PCMC we will require a better bound on MCMC than that given in Herbster et al. (2016) - one that can adapt to the number of rows and columns that actually take part in it. To describe our mistake bound for MCMC we will need to define, for all $L \subseteq [T]$, $\theta \in [0, 1]$ and $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$, the following sets:

$$A(L) := \{i \in [m] \mid \exists\, t \in L \,:\, i_t = i\} \quad ; \quad B(L) := \{j \in [n] \mid \exists\, t \in L \,:\, j_t = j\}$$

$$M(L) := \{t \in L \mid \hat{y}_t \neq y_t\} \quad ; \quad Z(L, \theta, \boldsymbol{P}, \boldsymbol{Q}) := \{t \in M(L) \mid |\boldsymbol{p}_{i_t} \cdot \boldsymbol{q}_{j_t}| < \theta\}\,.$$

Our bound for MCMC is then given by the following theorem:

Theorem 11 *For all* $(\ell, \dagger) \in [d] \times \{\circ, \bullet\}$ *and all* $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ *we have:*

$$|M(L_\ell^\dagger)| \leq \frac{\ln(m+n)}{c\theta_\ell^2}(|A(L_\ell^\dagger)| + |B(L_\ell^\dagger)|) + \frac{1}{c}|Z(L_\ell^\dagger, \theta_\ell, \boldsymbol{P}, \boldsymbol{Q})|\,.$$

Since the total number of mistakes of PCMC is the sum of $|M(L_\ell^\dagger)|$ over all $(\ell, \dagger) \in [d] \times \{\circ, \bullet\}$, Theorem 11 serves as the starting point of our analysis (proof of Theorem 2).

8. Conclusion

We considered the problem of learning a binary-valued matrix online. We gave an algorithm for this problem which has a mistake bound almost as good as that of running the kernel perceptron algorithm separately for each row of the matrix when given the optimal kernel over the columns of the matrix - or as good as the same thing but with the transpose of the matrix. We showed how our bound improves over previous bounds for the same problem and gave examples of our bound on two natural models of matrix. A few open problems are:

- Moridomi et al. (2018) managed to remove (by modifying the algorithm) the factor $\mathcal{O}(\ln(m+n))$ appearing in the bound of Hazan et al. (2012) - can the same be done to our work?
- As for previous algorithms for this problem, our algorithm has a time complexity of $\mathcal{O}(m^3+n^3)$ per trial - can this be reduced?
- Herbster et al. (2020) extended the work of Herbster et al. (2016) by considering additional side information in the form of kernels over the rows and columns of the matrix - can the same be done with our work?

9. Acknowledgements

The author would like to thank Mark Herbster and Massimiliano Pontil for valuable discussions. This research was conducted in part whilst the author was employed by University College London. The Author was supported in part by the UK Defence Science and Technology Laboratory (Dstl) and Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/P009204/1.

References

Emmanuel J. Candès and Terence Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56:2053–2080, 2010.

Nicolò Cesa-Bianchi and Gábor Lugosi. Prediction, learning, and games. 2006.

Rina Foygel. Commentary on "near-optimal algorithms for online matrix prediction". In *COLT*, 2012.

Sally A. Goldman and Manfred K. Warmuth. Learning binary relations using weighted majority voting. *Machine Learning*, 20:245–271, 1993.

Sally A. Goldman, Ronald L. Rivest, and Robert E. Schapire. Learning binary relations and total orders. *30th Annual Symposium on Foundations of Computer Science*, pages 46–51, 1989.

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. *SIAM J. Comput.*, 46:744–773, 2012.

Mark Herbster and Massimiliano Pontil. Prediction on a graph with a perceptron. In *NIPS*, 2006.

Mark Herbster, Guy Lever, and Massimiliano Pontil. Online prediction on large diameter graphs. In *NIPS*, 2008.

Mark Herbster, Stephen Pasteris, and Fabio Vitale. Online sum-product computation over trees. In *NIPS*, 2012.

Mark Herbster, Stephen Pasteris, and Massimiliano Pontil. Mistake bounds for binary matrix completion. In *NIPS*, 2016.

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

Kenichiro Moridomi, Kohei Hatano, and Eiji Takimoto. Online linear optimization with the log-determinant regularizer. *IEICE Trans. Inf. Syst.*, 101-D:1511–1520, 2018.

Albert B. J. Novikoff. On convergence proofs for perceptrons. 1963.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *COLT*, 2013.

Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65 6:386–408, 1958.

Shai Shalev-Shwartz. Online learning and online convex optimization. *Found. Trends Mach. Learn.*, 4:107–194, 2012.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *COLT*, 2005.

Manfred K. Warmuth. Winnowing subspaces. In *ICML '07*, 2007.

Appendix A. Proofs

A.1. Proof of Theorem 1

Lemma 12 *For all $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ there exists $\boldsymbol{K} \in \mathcal{P}^n$ such that:*

$$\sum_{i\in[m]} \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2} \leq \sum_{i\in[m]} \max_{j\in[n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 .$$

Proof Without loss of generality assume that the vectors $\{\boldsymbol{q}_i \mid i \in [n]\}$ are linearly independent (otherwise take limits).

Let $\{\tilde{\boldsymbol{e}}_i \mid i \in [n]\}$ be an orthonormal basis for the span of $\{\boldsymbol{q}_i \mid i \in [n]\}$. Define the matrices $\tilde{\boldsymbol{P}} \in \mathbb{R}^{m\times n}$ and $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{n\times n}$ such that for $i, j \in [n]$ we have:

$$\tilde{q}_{i,j} := \boldsymbol{q}_i \cdot \tilde{\boldsymbol{e}}_j$$

and for all $(i,j) \in [m] \times [n]$

$$\tilde{p}_{i,j} := \boldsymbol{p}_i \cdot \tilde{\boldsymbol{e}}_j \,.$$

Since the term $1/(\boldsymbol{p}_i \cdot \boldsymbol{q}_j)^2$ decreases if $\boldsymbol{p}_i \cdot \boldsymbol{q}_j$ increases we can have, without loss of generality, that $\boldsymbol{p}_i$ is in the span of $\{\boldsymbol{q}_i \mid i \in [n]\}$ (otherwise we could remove the part perpendicular to the span and then normalise, not increasing the right hand side of the inequality in the lemma statement). This means that $\|\tilde{\boldsymbol{p}}_i\| = 1$. Note that we automatically have $\|\tilde{\boldsymbol{q}}_i\| = 1$.

Now define $\boldsymbol{K} := \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^\top$ noting that for all $i \in [n]$ we have $k_{i,i} = \tilde{\boldsymbol{q}}_i \cdot \tilde{\boldsymbol{q}}_i = \|\tilde{\boldsymbol{q}}_i\|^2 = 1$ so that $\rho(\boldsymbol{K})^2 = 1$. For all $i \in [m]$ define $\boldsymbol{w}_i$ such that $\boldsymbol{w}_i^\top = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{Q}}^{-1}$. Now for all $(i,j) \in [m] \times [n]$ we have:

$$\boldsymbol{w}_i \cdot \boldsymbol{k}_j = \boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{e}_j = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{Q}}^{-1}\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^\top \boldsymbol{e}_j = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{q}}_j = \boldsymbol{p}_i \cdot \boldsymbol{q}_j \tag{3}$$

so since $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ we have that $\operatorname{sign}(\boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{e}_j) = \operatorname{sign}(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) = u_{i,j}$. This means that $\operatorname{sign}(\boldsymbol{K}\boldsymbol{w}_i) = \boldsymbol{u}_i$. We also have:

$$\boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{w}_i = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{Q}}^{-1}\tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^\top (\tilde{\boldsymbol{Q}}^{-1})^\top \tilde{\boldsymbol{p}}_i = \tilde{\boldsymbol{p}}_i^\top \boldsymbol{I}\boldsymbol{I}\tilde{\boldsymbol{p}}_i = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{p}}_i = \|\tilde{\boldsymbol{p}}_i\|^2 = 1$$

and hence $\boldsymbol{w}_i \in \Lambda(\boldsymbol{K}, \boldsymbol{u}_i)$. By Equation (3) we then have:

$$\delta(\boldsymbol{K}, \boldsymbol{u}_i) \geq \min_{j \in [n]} |\boldsymbol{w}_i \cdot \boldsymbol{k}_j| = \min_{j \in [n]} |\boldsymbol{p}_i \cdot \boldsymbol{q}_j|$$

so:

$$\sum_{i \in [m]} \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2} \leq \sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2$$

as required. ■

Lemma 13 *For all $\boldsymbol{K} \in \mathcal{P}^n$ there exists $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ such that:*

$$\sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 \leq \sum_{i \in [m]} \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2} \,.$$

Proof Since $\boldsymbol{K}$ is positive semi-definite there exists $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{n \times n}$ such that $\boldsymbol{K} = \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^\top$. Given such a $\tilde{\boldsymbol{Q}}$ let $\boldsymbol{Q}$ be such that for all $i \in [n]$ we have $\boldsymbol{q}_i = \tilde{\boldsymbol{q}}_i / \|\tilde{\boldsymbol{q}}_i\|$ noting that $\|\boldsymbol{q}_i\| = 1$ as required. Construct $\tilde{\boldsymbol{P}} \in \mathbb{R}^{m \times n}$ as follows. For all $i \in [m]$ let:

$$\boldsymbol{w}_i := \operatorname{argmax}_{\boldsymbol{w} \in \Lambda(\boldsymbol{K}, \boldsymbol{u}_i)} \min_{j \in [n]} |\boldsymbol{w} \cdot \boldsymbol{k}_j|$$

and let $\tilde{\boldsymbol{p}}_i := \tilde{\boldsymbol{Q}}^\top \boldsymbol{w}_i$. Then define $\boldsymbol{P} \in \mathbb{R}^{m \times n}$ such that for all $i \in [m]$ we have $\boldsymbol{p}_i = \tilde{\boldsymbol{p}}_i / \|\tilde{\boldsymbol{p}}_i\|$ noting that $\|\boldsymbol{p}_i\| = 1$ as required. For all $(i,j) \in [m] \times [n]$ we now have, by definition of $\boldsymbol{w}_i$, that:

$$|\tilde{\boldsymbol{p}}_i \cdot \tilde{\boldsymbol{q}}_j| = |\tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{q}}_j| = |(\tilde{\boldsymbol{Q}}^\top \boldsymbol{w}_i)^\top (\tilde{\boldsymbol{Q}}^\top \boldsymbol{e}_j)| = |\boldsymbol{w}_i^\top \tilde{\boldsymbol{Q}}\tilde{\boldsymbol{Q}}^\top \boldsymbol{e}_j| = |\boldsymbol{w}_i^\top \boldsymbol{K}\boldsymbol{e}_j| = |\boldsymbol{w} \cdot \boldsymbol{k}_j| \geq \delta(\boldsymbol{K}, \boldsymbol{u}_i)$$

and hence:

$$\left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 \leq \frac{\|\tilde{\boldsymbol{p}}_i\|^2 \|\tilde{\boldsymbol{q}}_j\|^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2} \,.$$

Note also that since $\boldsymbol{w}_i \in \Lambda(\boldsymbol{K}, \boldsymbol{u}_i)$ we have:

$$\|\tilde{\boldsymbol{p}}_i\|^2 = \tilde{\boldsymbol{p}}_i^\top \tilde{\boldsymbol{p}}_i = \boldsymbol{w}_i^\top \tilde{\boldsymbol{Q}} \tilde{\boldsymbol{Q}}^\top \boldsymbol{w}_i = \boldsymbol{w}_i^\top \boldsymbol{K} \boldsymbol{w}_i = 1$$

and we have:

$$\|\tilde{\boldsymbol{q}}_j\|^2 = \tilde{\boldsymbol{q}}_j^\top \tilde{\boldsymbol{q}}_j = \boldsymbol{e}_j^\top \tilde{\boldsymbol{Q}} \tilde{\boldsymbol{Q}}^\top \boldsymbol{e}_j = \boldsymbol{e}_j^\top \boldsymbol{K} \boldsymbol{e}_j = k_{j,j} \leq \rho(\boldsymbol{K})^2$$

so that:

$$\left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 \leq \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2}$$

and hence:

$$\sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 \leq \sum_{i \in [m]} \frac{\rho(\boldsymbol{K})^2}{\delta(\boldsymbol{K}, \boldsymbol{u}_i)^2}$$

as required. ■

By the definition of $\tau(\boldsymbol{U})$, Lemma 12 implies that:

$$\tau(\boldsymbol{U}) \leq \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})} \sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2$$

whilst Lemma 13, implies that:

$$\tau(\boldsymbol{U}) \geq \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})} \sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 .$$

■

A.2. Proof of Theorem 2

In this proof we adopt the notation defined in Section 7 and the PCMC algorithm. By Theorem 1 choose $(\boldsymbol{P}^\circ, \boldsymbol{Q}^\circ) \in \mathcal{N}(\boldsymbol{U})$ and $(\boldsymbol{P}^\bullet, \boldsymbol{Q}^\bullet) \in \mathcal{N}(\boldsymbol{U})$ such that:

$$\tau(\boldsymbol{U}) = \sum_{i \in [m]} \max_{j \in [n]} \frac{1}{(\boldsymbol{p}_i^\circ \cdot \boldsymbol{q}_j^\circ)^2} \quad ; \quad \tau(\boldsymbol{U}^\top) = \sum_{j \in [n]} \max_{i \in [m]} \frac{1}{(\boldsymbol{p}_i^\bullet \cdot \boldsymbol{q}_j^\bullet)^2} \tag{4}$$

and for all $(i, j) \in [m] \times [n]$ define:

$$\alpha_i := \min_{j \in [n]} |\boldsymbol{p}_i^\circ \cdot \boldsymbol{q}_j^\circ| \quad ; \quad \beta_j := \min_{i \in [m]} |\boldsymbol{p}_i^\bullet \cdot \boldsymbol{q}_j^\bullet| .$$

For all $\ell \in \{0\} \cup [d]$ define:

$$\Psi_\ell := \{i \in [m] \mid \alpha_i \leq \theta_\ell\} \quad ; \quad \Omega_\ell := \{j \in [n] \mid \beta_j \leq \theta_\ell\} .$$

For all $(i, j) \in [m] \times [n]$ let $\tilde{\lambda}_i^\circ$ and $\tilde{\lambda}_j^\bullet$ be the values of λ_i° and $\lambda_j^\bullet$ on trial T respectively. For all $\ell \in \{0\} \cup [d]$ we then define:

$$\tilde{\Psi}_\ell := \{i \in [m] \mid \tilde{\lambda}_i^\circ > \ell\} \quad ; \quad \tilde{\Omega}_\ell := \{j \in [n] \mid \tilde{\lambda}_j^\bullet > \ell\} .$$

Given $\ell \in [d]$ and some $i \in [m]$ we now bound the number of trials $t \in M(L_\ell^\circ)$ in which $i_t = i$. If $t \in M(L_\ell^\circ)$ and $i_t = i$ then $t \in L_\ell^\circ$ so we have $\ell_t = \ell$ and $\dagger_t = \circ$ and hence we must also have $\lambda_i^\circ = \lambda_{i_t}^\circ = \ell$ at the start of trial t, and have $\kappa_t = i_t = i$. Also, since $t \in M(L_\ell^\circ)$ we have $\hat{y}_t \neq u_{i_t,j_t}$ and hence $\mu_{\kappa_t}^{\dagger_t} \equiv \mu_i^\circ$ increases by one or $\lambda_{\kappa_t}^{\dagger_t} \equiv \lambda_i^\circ$ increases by one on trial t. Since λ_i° increases by one if and only if $\mu_i^\circ = \pi_{\ell_t} \equiv \pi_\ell$, and that when λ_i° increases we never have $\lambda_i^\circ = \ell$ again, there can only be at most π_ℓ such trials. In addition, there must be exactly π_ℓ such trials if there exists a trial on which $\lambda_i^\circ > \ell$. Since $i \in \tilde{\Psi}_\ell$ if and only if there exists a trial with $\lambda_i^\circ > \ell$ we then have that for all $i \in \tilde{\Psi}_\ell$ there exists π_ℓ trials $t \in M(L_\ell^\circ)$ with $i_t = i$. Summing over all $i \in \tilde{\Psi}_\ell$ gives us $|M(L_\ell^\circ)| \geq \pi_\ell |\tilde{\Psi}_\ell|$.

By applying this argument with $j \in [n]$ instead of i and $\bullet$ instead of $\circ$ we then have the following inequalities:

$$\forall (\ell, i) \in [d] \times [m]\,, \sum_{t \in M(L_\ell^\circ)} [\![i_t = i]\!] \leq \pi_\ell \quad ; \quad \forall (\ell, j) \in [d] \times [n]\,, \sum_{t \in M(L_\ell^\bullet)} [\![j_t = j]\!] \leq \pi_\ell \tag{5}$$

$$\forall \ell \in [d]\,,\ |M(L_\ell^\circ)| \geq \pi_\ell |\tilde{\Psi}_\ell| \quad ; \quad \forall \ell \in [d]\,,\ |M(L_\ell^\bullet)| \geq \pi_\ell |\tilde{\Omega}_\ell|\,. \tag{6}$$

Given $\ell \in [d]$ we now bound $|A(L_\ell^\circ)| + |B(L_\ell^\circ)|$ as follows. Choose any $i \in A(L_\ell^\circ)$. Then there exists a trial $t \in L_\ell^\circ$ with $i_t = i$. Since $t \in L_\ell^\circ$ we have $\dagger_t = \circ$ and $\ell_t = \ell$. We must then have $\lambda_i^\circ = \lambda_{i_t}^\circ = \ell > \ell - 1$ at the start of trial t. This implies that $i \in \tilde{\Psi}_{\ell-1}$. Now choose any $j \in B(L_\ell^\circ)$. Then there exists a trial $t \in L_\ell^\circ$ with $j_t = j$. Since $t \in L_\ell^\circ$ we have $\dagger_t = \circ$ and $\ell_t = \ell$. We must then have $\lambda_{i_t}^\circ = \ell > \ell - 1$ and $\lambda_j^\bullet = \lambda_{j_t}^\bullet \geq \lambda_{i_t}^\circ$ so $\lambda_j^\bullet > \ell - 1$ at the start of trial t. This implies that $j \in \tilde{\Omega}_{\ell-1}$. So we have shown that $A(L_\ell^\circ) \subseteq \tilde{\Psi}_{\ell-1}$ and $B(L_\ell^\circ) \subseteq \tilde{\Omega}_{\ell-1}$.

By applying the same argument to $A(L_\ell^\bullet)$ and $B(L_\ell^\bullet)$ we then have the following inequality.

$$\forall (\ell, \dagger) \in [d] \times \{\circ, \bullet\}\,,\ |A(L_\ell^\dagger)| + |B(L_\ell^\dagger)| \leq |\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|\,. \tag{7}$$

Given $\ell \in [d]$ we can now bound $|M(L_\ell^\circ)|$ as follows. Given $t \in Z(L_\ell^\circ, \theta_\ell, \boldsymbol{P}^\circ, \boldsymbol{Q}^\circ)$ we must have that $t \in M(L_\ell^\circ)$ and that $|\boldsymbol{p}_{i_t}^\circ \cdot \boldsymbol{q}_{j_t}^\circ| < \theta_\ell$ which implies that $i_t \in \Psi_\ell$. Hence we have:

$$Z(L_\ell^\circ, \theta_\ell, \boldsymbol{P}^\circ, \boldsymbol{Q}^\circ) \subseteq \{t \in M(L_\ell^\circ) \mid i_t \in \Psi_\ell\}$$

so by Equation (5) we have:

$$|Z(L_\ell^\circ, \theta_\ell, \boldsymbol{P}^\circ, \boldsymbol{Q}^\circ)| \leq \sum_{t \in M(L_\ell^\circ)} [\![i_t \in \Psi_\ell]\!] = \sum_{i \in \Psi_\ell} \sum_{t \in M(L_\ell^\circ)} [\![i_t = i]\!] \leq \sum_{i \in \Psi_\ell} \pi_\ell = \pi_\ell |\Psi_\ell|\,.$$

By combining this inequality with Equation (7) and Theorem 11 with $(\boldsymbol{P}, \boldsymbol{Q}) := (\boldsymbol{P}^\circ, \boldsymbol{Q}^\circ)$ we then have:

$$|M(L_\ell^\circ)| \leq \frac{\ln(m+n)}{c\theta_\ell^2}(|\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|) + \frac{1}{c}\pi_\ell |\Psi_\ell| = \frac{\pi_\ell}{2g}(|\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|) + \frac{1}{c}\pi_\ell |\Psi_\ell|\,. \tag{8}$$

By the same argument for $\bullet$ instead of $\circ$ we also have:

$$|M(L_\ell^\bullet)| \leq \frac{\pi_\ell}{2g}(|\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|) + \frac{1}{c}\pi_\ell |\Omega_\ell|\,. \tag{9}$$

Combining equations 8 and 9 gives us:

$$|M(L_\ell^\circ)| + |M(L_\ell^\bullet)| \leq \frac{\pi_\ell}{g}(|\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|) + \frac{\pi_\ell}{c}(|\Psi_\ell| + |\Omega_\ell|) \tag{10}$$

so that Equation 6 gives us:

$$|\tilde{\Psi}_\ell| + |\tilde{\Omega}_\ell| \leq \frac{1}{\pi_\ell}(|M(L_\ell^\circ)| + |M(L_\ell^\bullet)|) \leq \frac{1}{g}(|\tilde{\Omega}_{\ell-1}| + |\tilde{\Psi}_{\ell-1}|) + \frac{1}{c}(|\Psi_\ell| + |\Omega_\ell|)\,. \tag{11}$$

Since $\tilde{\Psi}_0 = [m] = \Psi_0$ and $\tilde{\Omega}_0 = [n] = \Omega_0$ we have:

$$|\tilde{\Psi}_0| + |\tilde{\Omega}_0| = |\Psi_0| + |\Omega_0| < \frac{1}{c}(|\Psi_0| + |\Omega_0|)\,. \tag{12}$$

The following inequality, which holds for all $\ell \in [d]$, is then easily verified via induction on ℓ using equations 11 and 12.

$$|\tilde{\Psi}_{\ell-1}| + |\tilde{\Omega}_{\ell-1}| \leq \frac{1}{c}\sum_{k=0}^{\ell-1} \frac{1}{g^{(\ell-1-k)}}(|\Psi_k| + |\Omega_k|)\,.$$

Substituting this inequality back into Equation 10 gives us:

$$|M(L_\ell^\circ)| + |M(L_\ell^\bullet)| \leq \frac{\pi_\ell}{c}\sum_{k=0}^{\ell} \frac{1}{g^{(\ell-k)}}(|\Psi_k| + |\Omega_k|)$$

so:

$$\begin{aligned}
\mathcal{M} &= \sum_{\ell=1}^{d}(|M(L_\ell^\circ)| + |M(L_\ell^\bullet)|) \leq \frac{1}{c}\sum_{\ell=1}^{d}\sum_{k=0}^{\ell} \frac{\pi_\ell}{g^{(\ell-k)}}(|\Psi_k| + |\Omega_k|) \\
&\leq \frac{1}{c}\sum_{k=0}^{d}(|\Psi_k| + |\Omega_k|)\sum_{\ell=k}^{d} \frac{\pi_\ell}{g^{(\ell-k)}} = \frac{2g\ln(m+n)}{c^2}\sum_{k=0}^{d}(|\Psi_k| + |\Omega_k|)\sum_{\ell=k}^{d} \frac{f^\ell}{g^{(\ell-k)}} \\
&= \frac{2g\ln(m+n)}{c^2}\sum_{k=0}^{d}(|\Psi_k| + |\Omega_k|)f^k \sum_{\ell=k}^{d}\left(\frac{f}{g}\right)^{\ell-k} \\
&\leq \frac{2g\ln(m+n)}{c^2}\sum_{k=0}^{d}(|\Psi_k| + |\Omega_k|)f^k \sum_{s=0}^{\infty}\left(\frac{f}{g}\right)^{s} = \frac{2g\ln(m+n)}{c^2(1-f/g)}\sum_{k=0}^{d}(|\Psi_k| + |\Omega_k|)f^k
\end{aligned} \tag{13}$$

since $f < g$. Now, given $i \in [m]$ let $\zeta_i := \lfloor \log_f(1/\alpha_i^2) \rfloor$. Note that for all $\ell \in [d]$ with $\ell > \zeta_i$ we have, since $f > 1$, that:

$$\theta_\ell^2 = (1/f)^\ell \leq (1/f)^{\zeta_i+1} < (1/f)^{\log_f(1/\alpha_i^2)} = \alpha_i^2$$

so $i \notin \Psi_\ell$. This gives us:

$$\sum_{k=0}^{d}[\![i \in \Psi_k]\!]f^k \leq \sum_{k=0}^{\zeta_i} f^k = f^{\zeta_i}\sum_{k=0}^{\zeta_i} f^{k-\zeta_i} \leq f^{\zeta_i}\sum_{s=0}^{\infty} f^{-s} = \frac{f^{\zeta_i}}{1-1/f} \leq \frac{1}{(1-1/f)\alpha_i^2}$$

which implies, by Equation (4) and definition of α_i:

$$\sum_{k=0}^{d} |\Psi_k| f^k = \sum_{k=0}^{d} \sum_{i\in[m]} [\![i \in \Psi_k]\!] f^k = \sum_{i\in[m]} \sum_{k=0}^{d} [\![i \in \Psi_k]\!] f^k \leq \frac{1}{1-1/f} \sum_{i\in[m]} \frac{1}{\alpha_i^2} = \frac{\tau(\boldsymbol{U})}{1-1/f}\,.$$

Using also the same argument, with $j \in [n]$ instead of i, we now have:

$$\sum_{k=0}^{d} |\Psi_k| f^k \leq \frac{\tau(\boldsymbol{U})}{1-1/f} \quad ; \quad \sum_{k=0}^{d} |\Omega_k| f^k \leq \frac{\tau(\boldsymbol{U}^\top)}{1-1/f}\,.$$

Substituting these inequalities into Equation (13) then gives us:

$$\mathcal{M} \leq \frac{2g\ln(m+n)}{c^2(1-f/g)(1-1/f)}(\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))$$

which upon setting $f := 2$, $g := 4$ and $c := 3 - e$ gives us:

$$\mathcal{M} \leq \frac{32}{(3-e)^2}\ln(m+n)(\tau(\boldsymbol{U}) + \tau(\boldsymbol{U}^\top))\,.$$

■

A.3. Proof of Theorem 3

Note first that for all $t \in [T]$ we have, since $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$, that $y_t \tilde{u}_{i_t,j_t} = |\tilde{u}_{i_t,j_t}| \geq \tilde{u}_* = 1/g$. This implies that

$$\ell_t(\tilde{u}_{i_t,j_t}) = \max\{0, -g y_t \tilde{u}_{i_t,j_t} + 1\} \leq \max\left\{0, -\frac{g}{g} + 1\right\} = 0$$

so since $\ell_t(\tilde{u}_{i_t,j_t}) \geq 0$ we must have $\ell_t(\tilde{u}_{i_t,j_t}) = 0$ and hence:

$$\sum_{t\in Z} \ell_t(\tilde{u}_{i_t,j_t}) = 0\,. \tag{14}$$

Now note that if $t \in Z$ then $y_t \neq \operatorname{sign}(w_t)$ so $y_t w_t \leq 0$ which implies:

$$\ell_t(w_t) = \max\{0, -g y_t w_t + 1\} \geq \max\{0, 1\} = 1$$

so we have:

$$\sum_{t\in Z} \ell_t(w_t) \geq \sum_{t\in Z} 1 = |Z|\,. \tag{15}$$

Substituting equations (14) and (15) into Equation (1) gives us:

$$|Z| \leq 2g \min_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{D}(\tilde{\boldsymbol{U}})} \sqrt{\tilde{\tau}(\boldsymbol{A},\boldsymbol{B})\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})\ln(2(m+n))|Z|}\,.$$

Squaring both sides an dividing through by $|Z|$ then gives us:

$$|Z| = 4g^2 \min_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{D}(\tilde{\boldsymbol{U}})} \tilde{\tau}(\boldsymbol{A},\boldsymbol{B})\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})\ln(2(m+n))\,.$$

Then substituting in $|Z| = \sum_{t\in[T]} [\![y_t \neq \operatorname{sign}(w_t)]\!]$ and $g = 1/\tilde{u}_*$ gives us the result. ■

A.4. Proof of Theorem 4

Lemma 14 *For all $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times n}$ we have:*

$$\min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i\in[m]} \|\boldsymbol{p}_i\|^2 \max_{j\in[n]} \|\boldsymbol{q}_j\|^2 + \sum_{j\in[n]} \|\boldsymbol{q}_j\|^2 \max_{i\in[m]} \|\boldsymbol{p}_i\|^2 \right) \leq 2 \min_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{D}(\tilde{\boldsymbol{U}})} \tilde{\tau}(\boldsymbol{A},\boldsymbol{B})\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})\,.$$

Proof Choose the pair $(\boldsymbol{A},\boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})$ that minimises $\tilde{\tau}(\boldsymbol{A},\boldsymbol{B})\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})$. Since $\boldsymbol{A}$ is positive semi-definite we can find some $\boldsymbol{C} \in \mathbb{R}^{(m+n)\times(m+n)}$ such that $\boldsymbol{A} = \boldsymbol{C}\boldsymbol{C}^\top$ so choose such a $\boldsymbol{C}$. Now choose $\boldsymbol{P}^+ \in \mathbb{R}^{m\times(m+n)}$ and $\boldsymbol{Q}^+ \in \mathbb{R}^{n\times(m+n)}$ such that $[\boldsymbol{P}^+,\boldsymbol{Q}^+] = \boldsymbol{C}$. We now have the following equalities:

$$\forall (i,j) \in [m]\times[n]\,,\;\; a_{i,(m+j)} = \boldsymbol{c}_i \cdot \boldsymbol{c}_{(m+j)} = \boldsymbol{p}_i^+ \cdot \boldsymbol{q}_j^+\,.$$

$$\max_{i\in[m+n]} a_{i,i} = \max_{i\in[m+n]} \boldsymbol{c}_i \cdot \boldsymbol{c}_i = \max_{i\in[m+n]} \|\boldsymbol{c}_i\|^2 = \max\left\{ \max_{i\in[m]} \|\boldsymbol{p}_i^+\|^2, \max_{i\in[n]} \|\boldsymbol{q}_i^+\|^2 \right\}\,.$$

$$\mathrm{Tr}(\boldsymbol{A}) = \sum_{i\in[m+n]} a_{i,i} = \sum_{i\in[m+n]} \boldsymbol{c}_i \cdot \boldsymbol{c}_i = \sum_{i\in[m+n]} \|\boldsymbol{c}_i\|^2 = \sum_{i\in[m]} \|\boldsymbol{p}_i^+\|^2 + \sum_{i\in[n]} \|\boldsymbol{q}_i^+\|^2\,.$$

Similarly there exist $\boldsymbol{P}^- \in \mathbb{R}^{m\times(m+n)}$ and $\boldsymbol{Q}^- \in \mathbb{R}^{n\times(m+n)}$ such that:

$$\forall (i,j) \in [m]\times[n],\; b_{i,(m+j)} = \boldsymbol{p}_i^- \cdot \boldsymbol{q}_j^-\,.$$

$$\max_{i\in[m+n]} b_{i,i} = \max\left\{ \max_{i\in[m]} \|\boldsymbol{p}_i^-\|^2, \max_{i\in[n]} \|\boldsymbol{q}_i^-\|^2 \right\}\,.$$

$$\mathrm{Tr}(\boldsymbol{B}) = \sum_{i\in[m]} \|\boldsymbol{p}_i^-\|^2 + \sum_{i\in[n]} \|\boldsymbol{q}_i^-\|^2\,.$$

Now define $\boldsymbol{P}$ and $\boldsymbol{Q}$ such that:

$$\boldsymbol{P}^\top := [(\boldsymbol{P}^+)^\top, -(\boldsymbol{P}^-)^\top]$$

and:

$$\boldsymbol{Q}^\top := [(\boldsymbol{Q}^+)^\top, (\boldsymbol{Q}^-)^\top]\,.$$

Since $(\boldsymbol{A},\boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})$ we have $\boldsymbol{A} - \boldsymbol{B} = \mathrm{sym}(\tilde{\boldsymbol{U}})$ so for all $(i,j) \in [m]\times[n]$ we have

$$a_{i,(m+j)} - b_{i,(m+j)} = \tilde{u}_{i,j}$$

and hence, by the above equalities and definitions of $\boldsymbol{P}$ and $\boldsymbol{Q}$, we have:

$$\boldsymbol{p}_i \cdot \boldsymbol{q}_j = \boldsymbol{p}_i^+ \cdot \boldsymbol{q}_j^+ - \boldsymbol{p}_i^- \cdot \boldsymbol{q}_j^- = a_{i,(m+j)} - b_{i,(m+j)} = \tilde{u}_{i,j}$$

so $(\boldsymbol{P},\boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$.

Now, by the above equalities and definition of $\boldsymbol{P}$ we have:

$$\max_{i\in[m]} \|\boldsymbol{p}_i\|^2 = \max_{i\in[m]} \left(\|\boldsymbol{p}_i^+\|^2 + \|\boldsymbol{p}_i^-\|^2 \right)$$

$$
\begin{aligned}
&\leq \max_{i\in[m]} \|\boldsymbol{p}_i^+\|^2 + \max_{i\in[m]} \|\boldsymbol{p}_i^-\|^2 \\
&\leq \max\left\{\max_{i\in[m]} \|\boldsymbol{p}_i^+\|^2, \max_{i\in[n]} \|\boldsymbol{q}_i^+\|^2\right\} + \max\left\{\max_{i\in[m]} \|\boldsymbol{p}_i^-\|^2, \max_{i\in[n]} \|\boldsymbol{q}_i^-\|^2\right\} \\
&= \max_{i\in[m+n]} a_{i,i} + \max_{i\in[m+n]} b_{i,i} \\
&\leq 2\max\left\{\max_{i\in[m+n]} a_{i,i}\,,\ \max_{i\in[m+n]} b_{i,i}\right\} \\
&= 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})
\end{aligned}
$$

and similarly:

$$\max_{j\in[n]} \|\boldsymbol{q}_j\|^2 \leq 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})$$

so, by the above equalities and definitions of $\boldsymbol{P}$ and $\boldsymbol{Q}$, we have:

$$
\begin{aligned}
&\sum_{i\in[m]} \|\boldsymbol{p}_i\|^2 \max_{j\in[n]} \|\boldsymbol{q}_j\|^2 + \sum_{j\in[n]} \|\boldsymbol{q}_j\|^2 \max_{i\in[m]} \|\boldsymbol{p}_i\|^2 \\
&\leq 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B}) \left(\sum_{i\in[m]} \|\boldsymbol{p}_i\|^2 + \sum_{j\in[n]} \|\boldsymbol{q}_j\|^2\right) \\
&= 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B}) \left(\sum_{i\in[m]} (\|\boldsymbol{p}_i^+\|^2 + \|\boldsymbol{p}_i^-\|^2) + \sum_{j\in[n]} (\|\boldsymbol{q}_j^+\|^2 + \|\boldsymbol{q}_j^-\|^2)\right) \\
&= 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B}) \left(\left(\sum_{i\in[m]} \|\boldsymbol{p}_i^+\|^2 + \sum_{i\in[n]} \|\boldsymbol{q}_i^+\|^2\right) + \left(\sum_{i\in[m]} \|\boldsymbol{p}_i^-\|^2 + \sum_{i\in[n]} \|\boldsymbol{q}_i^-\|^2\right)\right) \\
&= 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})(\mathrm{Tr}(\boldsymbol{A}) + \mathrm{Tr}(\boldsymbol{B})) \\
&= 2\tilde{\beta}(\boldsymbol{A},\boldsymbol{B})\tilde{\tau}(\boldsymbol{A},\boldsymbol{B})
\end{aligned}
$$

as required. ■

Lemma 15 *For all $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times n}$ we have:*

$$\min_{(\boldsymbol{A},\boldsymbol{B})\in\mathcal{D}(\tilde{\boldsymbol{U}})} \tilde{\tau}(\boldsymbol{A},\boldsymbol{B})\tilde{\beta}(\boldsymbol{A},\boldsymbol{B}) \leq 2 \min_{(\boldsymbol{P},\boldsymbol{Q})\in\mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i\in[m]} \|\boldsymbol{p}_i\|^2 \max_{j\in[n]} \|\boldsymbol{q}_j\|^2 + \sum_{j\in[n]} \|\boldsymbol{q}_j\|^2 \max_{i\in[m]} \|\boldsymbol{p}_i\|^2\right).$$

Proof Choose $(\boldsymbol{P},\boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$ that minimises the right hand side of this inequality. Note that for any number $x \in \mathbb{R}\setminus\{0\}$, multiplying $\boldsymbol{P}$ by x and dividing $\boldsymbol{Q}$ by x does not affect the term in brackets or the values $\boldsymbol{p}_i \cdot \boldsymbol{q}_j \equiv \tilde{u}_{i,j}$. This means that, without loss of generality, we may assume that:

$$\max_{i\in[m]} \|\boldsymbol{p}_i\|^2 = \max_{j\in[n]} \|\boldsymbol{q}_j\|^2\,. \tag{16}$$

Let $\boldsymbol{C} := [\boldsymbol{P},\boldsymbol{Q}]$ and define $\boldsymbol{A} := \boldsymbol{C}\boldsymbol{C}^\top$. Then define:

$$\boldsymbol{B} := [[\boldsymbol{P}\boldsymbol{P}^\top, \boldsymbol{0}^{n,m}]^\top, [\boldsymbol{0}^{m,n}, \boldsymbol{Q}\boldsymbol{Q}^\top]^\top]\,.$$

For all $i, j \in [m]$ we have $a_{i,j} = \boldsymbol{p}_i \cdot \boldsymbol{p}_j = b_{i,j}$ so $a_{i,j} - b_{i,j} = 0$. Similarly for all $i, j \in [n]$ we have $a_{(m+i),(m+j)} = \boldsymbol{q}_i \cdot \boldsymbol{q}_j = b_{(m+i),(m+j)}$ so $a_{(m+i),(m+j)} - b_{(m+i),(m+j)} = 0$. Also, for all $(i, j) \in [m] \times [n]$ we have $a_{i,(m+j)} = \boldsymbol{p}_i \cdot \boldsymbol{q}_j = \tilde{u}_{i,j}$ and $b_{i,(m+j)} = 0$ so $a_{i,(m+j)} - b_{i,(m+j)} = \tilde{u}_{i,j}$ and similarly we have $a_{(m+j),i} - b_{(m+j),i} = \tilde{u}_{j,i}$. This implies that $\boldsymbol{A} - \boldsymbol{B} = \operatorname{sym}(\tilde{\boldsymbol{U}})$ so since both $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive semidefinite we have $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})$.

Note that for all $i, j \in [m]$ we have $a_{i,i} = \boldsymbol{p}_i \cdot \boldsymbol{p}_i$ and $a_{(m+j),(m+j)} = \boldsymbol{q}_j \cdot \boldsymbol{q}_j$ so:

$$\sum_{l \in [m+n]} a_{l,l} = \sum_{i \in [m]} a_{i,i} + \sum_{j \in [n]} a_{(m+j),(m+j)} = \sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2$$

and:

$$\max_{l \in [m+n]} a_{l,l} = \max \left\{ \max_{i \in [m]} a_{i,i} \,,\, \max_{j \in [n]} a_{(m+j),(m+j)} \right\} = \max \left\{ \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 \,,\, \max_{j \in [n]} \|\boldsymbol{q}_i\|^2 \right\}$$

which by Equation (16) gives us:

$$\max_{l \in [m+n]} a_{l,l} = \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 = \max_{j \in [n]} \|\boldsymbol{q}_i\|^2 .$$

Similarly we have:

$$\sum_{l \in [m+n]} b_{l,l} = \sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \quad \text{and} \quad \max_{l \in [m+n]} b_{l,l} = \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 = \max_{j \in [n]} \|\boldsymbol{q}_i\|^2$$

so:

$$\tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) = 2 \left(\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \right) \quad \text{and} \quad \tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) = \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 = \max_{j \in [n]} \|\boldsymbol{q}_i\|^2$$

which gives us:

$$\tilde{\tau}(\boldsymbol{A}, \boldsymbol{B}) \tilde{\beta}(\boldsymbol{A}, \boldsymbol{B}) = 2 \left(\sum_{i \in [m]} \|\boldsymbol{p}_i\|^2 \max_{j \in [n]} \|\boldsymbol{q}_j\|^2 + \sum_{j \in [n]} \|\boldsymbol{q}_j\|^2 \max_{i \in [m]} \|\boldsymbol{p}_i\|^2 \right)$$

which, since by above we have $(\boldsymbol{A}, \boldsymbol{B}) \in \mathcal{D}(\tilde{\boldsymbol{U}})$, implies the result. ■

Lemmas 14 and 15 imply the result. ■

A.5. Proof of Theorem 5

The result clearly holds for $k = 1$ so without loss of generality let $k \geq 2$. Let $f := 1/10$. We start with the following lemma.

Lemma 16 *There exists a matrix $\boldsymbol{R} \in \{-1, 1\}^{k \times k}$ such that for all $l \in \mathbb{N}$ and $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{k \times l}$ with $r_{i,j} \boldsymbol{p}_i \cdot \boldsymbol{q}_j \geq 1$ for all $i, j \in [k]$, we have:*

$$\max_{i,j \in [k]} \|\boldsymbol{p}_i\| \|\boldsymbol{q}_j\| \geq f \sqrt{k} .$$

Proof We use the (parameter-free) algorithm of Herbster et al. (2016), with k^2 trials, to build the matrix $\boldsymbol{R}$. Let (i'_t, j'_t) be the pair given to the algorithm on trial t. Given the algorithm's prediction $y'_t \in \{-1, 1\}$ we then set $r_{i'_t,j'_t} = -y'_t$. This defines $\boldsymbol{R}$. Now suppose we have $l \in \mathbb{N}$ and $\boldsymbol{P}, \boldsymbol{Q} \in \mathbb{R}^{k\times l}$ with $r_{i,j}\boldsymbol{p}_i \cdot \boldsymbol{q}_j \geq 1$ for all $i, j \in [k]$. Note that, by the mistake bound in Herbster et al. (2016), the number of mistakes made by the algorithm is no greater than $(k/f^2)\max_{i,j\in[k]} \|\boldsymbol{p}_i\|^2\|\boldsymbol{q}_j\|^2$. But the algorithm made a mistake on every trial and hence made k^2 mistakes. This means that $\max_{i,j\in[k]} \|\boldsymbol{p}_i\|^2\|\boldsymbol{q}_j\|^2 \geq f^2k$ so that $\max_{i,j\in[k]} \|\boldsymbol{p}_i\|\|\boldsymbol{q}_j\| \geq f\sqrt{k}$ as required. ■

Letting $m = n = k + k^2$, we define our matrix $\boldsymbol{U} \in \{-1, 1\}^{m\times n}$ as follows. Choose $\boldsymbol{R}$ so that it satisfies Lemma 16. For all $(i, j) \in [k]^2$ let $u_{i,j} = r_{i,j}$ and for all $(i, j) \in [m]^2 \setminus [k]^2$ let $u_{i,j} = 1 - 2[\![i \neq j]\!]$. We now have the following two lemmas:

Lemma 17 *We have:*

$$h(\boldsymbol{U}) \geq fk^{5/2}\,.$$

Proof Choose the minimising $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$ in the definition of $h(\boldsymbol{U})$. Since $h(\boldsymbol{U})$ is scale invariant, without loss of generality assume that $\tilde{u}_* = 1$. Next choose the minimising $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$ in the definition of $h(\boldsymbol{U})$. Note that for all $i, j \in [k]$ we have $u_{i,j}\boldsymbol{p}_i \cdot \boldsymbol{q}_j > 0$ so since $|\boldsymbol{p}_i \cdot \boldsymbol{q}_j| = |\tilde{u}_{i,j}| \geq 1$ and $u_{i,j} = r_{i,j}$ we have $r_{i,j}\boldsymbol{p}_i\cdot\boldsymbol{q}_j \geq 1$. So by Lemma 16 there exists $i, j \in [k]$ with $\|\boldsymbol{p}_i\|\|\boldsymbol{q}_j\| \geq f\sqrt{k}$. For all $i \in [m]$ we have $|\boldsymbol{p}_i \cdot \boldsymbol{q}_i| = |\tilde{u}_{i,i}| \geq 1$ which implies $\|\boldsymbol{p}_i\|\|\boldsymbol{q}_i\| \geq 1$. By above, this means that:

$$(\|\boldsymbol{p}_i\| \max_{l\in[m]} \|\boldsymbol{q}_l\|)(\|\boldsymbol{q}_i\| \max_{j\in[m]} \|\boldsymbol{p}_j\|) = (\|\boldsymbol{p}_i\|\|\boldsymbol{q}_i\|) \max_{j,l\in[m]} \|\boldsymbol{p}_j\|\|\boldsymbol{q}_l\| \geq (\|\boldsymbol{p}_i\|\|\boldsymbol{q}_i\|)f\sqrt{k} \geq f\sqrt{k}\,.$$

This means that either $\|\boldsymbol{p}_i\|^2 \max_{l\in[m]} \|\boldsymbol{q}_l\|^2 \geq f\sqrt{k}$ or $\|\boldsymbol{q}_i\|^2 \max_{j\in[m]} \|\boldsymbol{p}_j\|^2 \geq f\sqrt{k}$. Substituing into the definition of $h(\boldsymbol{U})$ gives us $h(\boldsymbol{U}) \geq mf\sqrt{k} \geq fk^{5/2}$. ■

Lemma 18 *We have:*

$$h'(\boldsymbol{U}) \leq 40k^2\,.$$

Proof We construct $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathbb{R}^{m\times(m+3)}$ as follows. For all $i \in [k]$ define $\boldsymbol{p}_i, \boldsymbol{q}_i \in \mathbb{R}^{m+3}$ as follows:

$$\forall j \in [k],\ (p_{i,j}, q_{i,j}) := (u_{i,j}, [\![i = j]\!]\sqrt{k})\,.$$

$$\forall j \in [m] \setminus [k],\ (p_{i,j}, q_{i,j}) := (0, 0)\,.$$

$$(p_{i,m+1}, q_{i,m+1}) := (\sqrt{k}, 0)\ ;\ (p_{i,m+2}, q_{i,m+2}) := (0, \sqrt{k})\ ;(p_{i,m+3}, q_{i,m+3}) := (0, 0)\,.$$

For all $i \in [m] \setminus [k]$ define $\boldsymbol{p}_i, \boldsymbol{q}_i \in \mathbb{R}^{m+3}$ as follows.

$$\forall j \in [k],\ (p_{i,j}, q_{i,j}) := (0, 0)\,.$$

$$\forall j \in [m] \setminus [k],\ (p_{i,j}, q_{i,j}) := (\sqrt{2}[\![i = j]\!], \sqrt{2}[\![i = j]\!])\,.$$

$$(p_{i,m+1}, q_{i,m+1}) := (0, -1)\ ;\ (p_{i,m+2}, q_{i,m+2}) := (-1, 0)\ ;\ (p_{i,m+3}, q_{i,m+3}) := (-1, 1)\,.$$

Note that for all $i \in [k]$ we have $\|\boldsymbol{p}_i\|^2 = \|\boldsymbol{q}_i\|^2 = 2k$ and for all $i \in [m] \setminus [k]$ we have $\|\boldsymbol{p}_i\|^2 = \|\boldsymbol{q}_i\|^2 = 4$. Now define $\tilde{\boldsymbol{U}} \in \mathbb{R}^{m\times m}$ such that for all $i, j \in [m]$ we have $\tilde{u}_{i,j} = \boldsymbol{p}_i \cdot \boldsymbol{q}_j$, noting that

$(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$. We then have the following identities. For all $i, j \in [k]$ we have $\tilde{u}_{i,j} = u_{i,j}\sqrt{k}$. For all $i, j \in [m] \setminus [k]$ we have $\tilde{u}_{i,j} = 2[\![i = j]\!] - 1$. For all $(i, j) \in [k] \times ([m] \setminus [k])$ we have $\tilde{u}_{i,j} = \tilde{u}_{j,i} = -\sqrt{k}$. So we have that $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$ and that:

$$\forall i \in [k]\,, \quad \max_{j \in [m]} \frac{\|\boldsymbol{q}_j\|^2}{(\boldsymbol{p}_i \cdot \boldsymbol{q}_j)^2} = \max_{j \in [m]} \frac{\|\boldsymbol{p}_j\|^2}{(\boldsymbol{q}_i \cdot \boldsymbol{p}_j)^2} = \max\left\{\frac{2k}{k}, \frac{4}{k}\right\} = 2$$

and:

$$\forall i \in [m] \setminus [k]\,, \quad \max_{j \in [m]} \frac{\|\boldsymbol{q}_j\|^2}{(\boldsymbol{p}_i \cdot \boldsymbol{q}_j)^2} = \max_{j \in [m]} \frac{\|\boldsymbol{p}_j\|^2}{(\boldsymbol{q}_i \cdot \boldsymbol{p}_j)^2} = \max\left\{\frac{2k}{k}, \frac{4}{1}\right\} = 4\,.$$

Substituting into the definition of $h'(\boldsymbol{U})$ we have:

$$h'(\boldsymbol{U}) \leq (k \cdot (2k) \cdot 2 + (m-k) \cdot 4 \cdot 4) + (k \cdot (2k) \cdot 2 + (m-k) \cdot 4 \cdot 4) = 40k^2$$

as required. ■

Lemmas 17 and 18 imply the result. ■

A.6. Proof of Theorem 6

Recall:

$$h'(\boldsymbol{U}) := \min_{\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})} \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})} \left(\sum_{i \in [m]} \max_{j \in [n]} \left(\frac{\|\boldsymbol{p}_i\| \|\boldsymbol{q}_j\|}{\tilde{u}_{i,j}} \right)^2 + \sum_{j \in [n]} \max_{i \in [m]} \left(\frac{\|\boldsymbol{p}_i\| \|\boldsymbol{q}_j\|}{\tilde{u}_{i,j}} \right)^2 \right).$$

Clearly:

$$h'(\boldsymbol{U}) \leq \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})} \left(\sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 + \sum_{j \in [n]} \max_{i \in [m]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 \right)$$

since if $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ then by defining $\tilde{\boldsymbol{U}} := \boldsymbol{P}\boldsymbol{Q}^\top$ we have $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$, $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$, and for all $(i, j) \in [m] \times [n]$ we have $\tilde{u}_{i,j} = \boldsymbol{p}_i \cdot \boldsymbol{q}_j$ and $\|\boldsymbol{p}_i\| = \|\boldsymbol{q}_j\| = 1$.

So all that's left is to prove is that:

$$h'(\boldsymbol{U}) \geq \min_{(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})} \left(\sum_{i \in [m]} \max_{j \in [n]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 + \sum_{j \in [n]} \max_{i \in [m]} \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 \right)$$

which is done by showing that for all $\tilde{\boldsymbol{U}} \in \mathcal{S}(\boldsymbol{U})$ and $(\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}}) \in \mathcal{F}(\tilde{\boldsymbol{U}})$ there exists $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ with:

$$\left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 = \left(\frac{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|}{\tilde{u}_{i,j}} \right)^2$$

for all $(i, j) \in [m] \times [n]$. To show this define, for all $(i, j) \in [m] \times [n]$, $\boldsymbol{p}_i := \tilde{\boldsymbol{p}}_i / \|\tilde{\boldsymbol{p}}_i\|$ and $\boldsymbol{q}_j := \tilde{\boldsymbol{q}}_j / \|\tilde{\boldsymbol{q}}_j\|$ so that:

$$\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} = \frac{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|}{\tilde{\boldsymbol{p}}_i \cdot \tilde{\boldsymbol{q}}_j} = \frac{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|}{\tilde{u}_{i,j}}\,.$$

■

A.7. Proof of Theorem 7

By definition of $\mathcal{E}(\boldsymbol{U})$ there exists $(\boldsymbol{P}', \boldsymbol{Q}') \in \mathcal{N}(\boldsymbol{U})$ and $(\boldsymbol{P}'', \boldsymbol{Q}'') \in \mathcal{N}(\boldsymbol{U})$ such that for all $(i, j) \in [m] \times [n]$ we have:

$$d_{i,j} \geq \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j} \right)^2 \quad \text{and} \quad f_{i,j} \geq \left(\frac{1}{\boldsymbol{p}''_i \cdot \boldsymbol{q}''_j} \right)^2 .$$

Now let $\boldsymbol{P}$ and $\boldsymbol{Q}$ be such that for all $(i, j) \in [m] \times [n]$ we have:

$$\boldsymbol{p}_i := \frac{1}{\sqrt{2}}[\boldsymbol{p}'_i \,,\, \boldsymbol{p}''_i] \quad \text{and} \quad \boldsymbol{q}_j := \frac{1}{\sqrt{2}}[\boldsymbol{q}'_j \,,\, \boldsymbol{q}''_j]$$

and let $\boldsymbol{C}$ be such that for all $(i, j) \in [m] \times [n]$ we have:

$$c_{i,j} := \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 .$$

Note that for all $(i, j) \in [m] \times [n]$ we have:

- $\|\boldsymbol{p}_i\|^2 = \frac{1}{2}(\|\boldsymbol{p}'_i\|^2 + \|\boldsymbol{p}''_i\|^2) = \frac{1}{2}(1 + 1) = 1\,.$
- $\|\boldsymbol{q}_j\|^2 = \frac{1}{2}(\|\boldsymbol{q}'_j\|^2 + \|\boldsymbol{q}''_j\|^2) = \frac{1}{2}(1 + 1) = 1\,.$
- $(\boldsymbol{p}_i \cdot \boldsymbol{q}_j)u_{i,j} = \frac{1}{2}((\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j)u_{i,j} + (\boldsymbol{p}''_i \cdot \boldsymbol{q}''_j)u_{i,j}) > \frac{1}{2}(0 + 0) = 0$

so $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ and hence $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$. Since $\operatorname{sign}(\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j) = \operatorname{sign}(\boldsymbol{p}''_i \cdot \boldsymbol{q}''_j)$ we also have:

$$\begin{aligned} c_{i,j} = \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 &= \left(\frac{1}{(\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j + \boldsymbol{p}''_i \cdot \boldsymbol{q}''_j)/2} \right)^2 = 4 \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j + \boldsymbol{p}''_i \cdot \boldsymbol{q}''_j} \right)^2 \\ &\leq 4 \min \left\{ \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j} \right)^2 , \left(\frac{1}{\boldsymbol{p}''_i \cdot \boldsymbol{q}''_j} \right)^2 \right\} \leq 4 \min\{d_{i,j}, f_{i,j}\} \end{aligned}$$

as required. ■

A.8. Proof of Theorem 8

Let $a := 1.5$ and let:

$$b := \left(1 + \frac{a^2}{4}\right)\left(1 + \frac{2}{a^2}\right)$$

noting that $b < 3$. By definition of $\mathcal{E}(\boldsymbol{U})$ there exists $(\boldsymbol{P}', \boldsymbol{Q}') \in \mathcal{N}(\boldsymbol{U})$ such that:

$$f_{i,j} \geq \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_j} \right)^2$$

for all $(i, j) \in [m] \times [n]$. For all $i \in [m]$ let:

$$z_i := \operatorname{argmin}_{z \in \{-1,1\}} \max_{j \in [n]} [\![u_{i,j} = z]\!] f_{i,j}$$

and let:

$$g_i := \frac{1}{2} \min_{j \in [n]: u_{i,j} = z_i} |\boldsymbol{p}_i' \cdot \boldsymbol{q}_j'|$$

where here the minimiser of the empty set is equal to 1. Note that since $\|\boldsymbol{p}_i'\| = \|\boldsymbol{q}_j'\| = 1$ we have $|\boldsymbol{p}_i' \cdot \boldsymbol{q}_j'| \leq 1$ for all $j \in [n]$ so $g_i \leq 1/2$. Now let $\boldsymbol{P}$ be such that for all $i \in [m]$ with $z_i = -1$ we have:

$$\boldsymbol{p}_i := \frac{1}{\sqrt{1 + a^2/4}} \left[\boldsymbol{p}_i', \left(a g_i, 0, \sqrt{a^2/4 - a^2 g_i^2} \right)^\top \right]$$

and for all $i \in [m]$ with $z_i = 1$ we have:

$$\boldsymbol{p}_i := \frac{1}{\sqrt{1 + a^2/4}} \left[\boldsymbol{p}_i', \left(0, a g_i, \sqrt{a^2/4 - a^2 g_i^2} \right)^\top \right].$$

Then let $\boldsymbol{Q}$ be such that for all $j \in [n]$ we have:

$$\boldsymbol{q}_j := \frac{1}{\sqrt{1 + 2/a^2}} \left[\boldsymbol{q}_j', \left(\frac{1}{a}, -\frac{1}{a}, 0 \right)^\top \right]$$

and let $\boldsymbol{C}$ be such that for all $(i, j) \in [m] \times [n]$ we have:

$$c_{i,j} := \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2.$$

Note first that for all $(i, j) \in [m] \times [n]$ with $z_i = -1$ we have:

$$\boldsymbol{p}_i \cdot \boldsymbol{q}_j = \frac{1}{\sqrt{b}} (\boldsymbol{p}_i' \cdot \boldsymbol{q}_j' + g_i). \tag{17}$$

If, in addition, $u_{i,j} = 1$ then $\operatorname{sign}(\boldsymbol{p}_i' \cdot \boldsymbol{q}_j') = u_{i,j} = 1$ so by Equation (17), we have that $\boldsymbol{p}_i \cdot \boldsymbol{q}_j \geq g_i/\sqrt{b}$. On the other hand, if $u_{i,j} = -1$ then we have that $\operatorname{sign}(\boldsymbol{p}_i' \cdot \boldsymbol{q}_j') = u_{i,j} = -1$ and, since $z_i = -1 = u_{i,j}$, we have:

$$|\boldsymbol{p}_i' \cdot \boldsymbol{q}_j'| \geq \min_{j \in [n]: u_{i,j} = z_i} |\boldsymbol{p}_i' \cdot \boldsymbol{q}_j'| = 2 g_i$$

so $\boldsymbol{p}_i' \cdot \boldsymbol{q}_j' \leq -2 g_i$ and hence by Equation (17) we have that $\boldsymbol{p}_i \cdot \boldsymbol{q}_j \leq -g_i/\sqrt{b}$. Hence, in either case, given $z_i = -1$ we have $(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) u_{i,j} \geq g_i/\sqrt{b}$. The same argument gives us $(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) u_{i,j} \geq g_i/\sqrt{b}$ when $z_i = 1$ as well. Hence we have, for all $(i, j) \in [m] \times [n]$, that:

$$(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) u_{i,j} \geq g_i/\sqrt{b}. \tag{18}$$

In addition, for all $(i, j) \in [m] \times [n]$ we have:

- $\|\boldsymbol{p}_i\|^2 = \frac{1}{1 + a^2/4} (\|\boldsymbol{p}_i'\|^2 + a^2/4) = \frac{1}{1 + a^2/4} (1 + a^2/4) = 1$
- $\|\boldsymbol{q}_j\|^2 = \frac{1}{1 + 2/a^2} (\|\boldsymbol{q}_j\|^2 + 2/a^2) = \frac{1}{1 + 2/a^2} (1 + 2/a^2) = 1$

which, combined with Equation 18, implies that $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$. Hence we have $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ as required. By Equation 18 we also have, for all $(i,j) \in [m] \times [n]$, that:

$$c_{i,j} = \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 \leq \frac{b}{g_i^2} = 4b \max_{j' \in [n] : u_{i,j'} = z_i} \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_{j'}}\right)^2$$

where here the maximiser of the empty set is equal to one. Noting that:

$$\max_{j' \in [n] : u_{i,j'} = z_i} \left(\frac{1}{\boldsymbol{p}'_i \cdot \boldsymbol{q}'_{j'}}\right)^2 = \min_{z \in \{-1,1\}} \max_{j' \in [n] : u_{i,j'} = z} f_{i,j'} \leq 1 + \min_{z \in \{-1,1\}} \max_{j' \in [n]} [\![u_{i,j'} = z]\!] f_{i,j'}$$

we have proved that there exists $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ with:

$$\max_{j \in [n]} c_{i,j} \leq 4b + 4b \min_{z \in \{-1,1\}} \max_{j' \in [n]} [\![u_{i,j'} = z]\!] f_{i,j'}$$

for all $i \in [m]$. By the same argument, but with $\boldsymbol{U}^\top$ instead of $\boldsymbol{U}$, we have that there also exists $\boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ with:

$$\max_{i \in [m]} d_{i,j} \leq 4b + 4b \min_{z \in \{-1,1\}} \max_{i' \in [m]} [\![u_{i',j} = z]\!] f_{i',j}$$

as required. ■

A.9. Proof of Theorem 9

For all $j \in [n]$ define:

$$\beta_j := \sum_{l \in [k]} [\![j \in B_l]\!] .$$

Define $\boldsymbol{Q} \in \mathbb{R}^{n \times (k+1)}$ by, for all $(j,l) \in [n] \times [k+1]$:

- If $l \in [k]$ then $q_{j,l} := \frac{1}{\sqrt{2\beta_j}} [\![j \in B_l]\!]$.
- If $l = k+1$ then $q_{j,l} := \frac{1}{\sqrt{2}}$.

Define $\tilde{\boldsymbol{P}} \in \mathbb{R}^{m \times (k+1)}$ by, for all $(i,l) \in [m] \times [k+1]$:

- If $l \in [k]$ then $\tilde{p}_{i,l} := [\![i \in A_l]\!] \max_{j \in B_l} \sqrt{\beta_j}$.
- If $l = k+1$ then $\tilde{p}_{i,l} := -1/2$

and then define $\boldsymbol{P} \in \mathbb{R}^{m \times (k+1)}$ by $\boldsymbol{p}_i := \tilde{\boldsymbol{p}}_i / \|\tilde{\boldsymbol{p}}_i\|$ for all $i \in [m]$, and define $\boldsymbol{C}$ such that for all $(i,j) \in [m] \times [n]$ we have:

$$c_{i,j} := \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 .$$

Given $(i,j) \in [m] \times [n]$ with $u_{i,j} = 1$ there exists $l \in [k]$ with $[\![i \in A_l]\!] [\![j \in B_l]\!] = 1$ so since $\tilde{p}_{i,l'} q_{i,l'} \geq 0$ for all $l' \in [k]$, we have:

$$\tilde{\boldsymbol{p}}_i \cdot \boldsymbol{q}_j = \tilde{p}_{i,k+1} q_{j,k+1} + \sum_{l' \in [k]} \tilde{p}_{i,l'} q_{j,l'}$$

$$
\begin{aligned}
&\geq \tilde{p}_{i,k+1} q_{j,k+1} + \tilde{p}_{i,l} q_{j,l} \\
&= -\frac{1}{2\sqrt{2}} + \left([\![i \in A_l]\!] \max_{j' \in B_l} \sqrt{\beta_{j'}} \right) \frac{1}{\sqrt{2\beta_j}} [\![j \in B_l]\!] \\
&= -\frac{1}{2\sqrt{2}} + \max_{j' \in B_l} \frac{\sqrt{\beta_{j'}}}{\sqrt{2\beta_j}} \\
&\geq -\frac{1}{2\sqrt{2}} + \frac{1}{\sqrt{2}} \\
&= \frac{1}{2\sqrt{2}} \, .
\end{aligned}
$$

Given $(i, j) \in [m] \times [n]$ with $u_{i,j} = -1$ we have $[\![i \in A_l]\!] [\![j \in B_l]\!] = 0$ and hence $\tilde{p}_{i,l} q_{j,l} = 0$ for all $l \in [k]$, which implies:

$$
\tilde{\boldsymbol{p}}_i \cdot \boldsymbol{q}_j = \tilde{p}_{i,k+1} q_{j,k+1} + \sum_{l \in [k]} \tilde{p}_{i,l} q_{j,l} = -\frac{1}{2\sqrt{2}} + 0 = -\frac{1}{2\sqrt{2}} \, .
$$

So we have proved, for all $(i, j) \in [m] \times [n]$, that $u_{i,j} \tilde{\boldsymbol{p}}_i \cdot \boldsymbol{q}_j \geq 1/(2\sqrt{2})$ and hence that:

$$
u_{i,j} \boldsymbol{p}_i \cdot \boldsymbol{q}_j \geq \frac{1}{2\sqrt{2} \|\tilde{\boldsymbol{p}}_i\|} \, . \tag{19}
$$

Now note that for all $j \in [n]$ we have:

$$
\|\boldsymbol{q}_j\|^2 = q_{j,k+1}^2 + \sum_{l \in [k]} q_{j,l}^2 = \frac{1}{2} + \sum_{l \in [k]} \frac{[\![j \in B_l]\!]}{2\beta_j} = \frac{1}{2} + \frac{\sum_{l \in [k]} [\![j \in B_l]\!]}{2\beta_j} = \frac{1}{2} + \frac{\beta_j}{2\beta_j} = 1
$$

so, combining with Equation (19) and the fact that $\|\boldsymbol{p}_i\| = 1$ for all $i \in [m]$, we have that $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ and hence that $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$.

For all $i \in [m]$ we have that:

$$
\|\tilde{\boldsymbol{p}}_i\|^2 = \tilde{p}_{i,k+1}^2 + \sum_{l \in [k]} \tilde{p}_{i,l}^2 = \frac{1}{4} + \sum_{l \in [k]} [\![i \in A_l]\!] \max_{j \in B_l} \beta_j = \frac{1}{4} + \sum_{l \in [k]} [\![i \in A_l]\!] \max_{j \in B_l} \sum_{l' \in [k]} [\![j \in B_{l'}]\!]
$$

so, by Equation (19), we have:

$$
c_{i,j} = \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j} \right)^2 \leq 8 \|\tilde{\boldsymbol{p}}_i\|^2 = 2 + 8 \sum_{l \in [k]} [\![i \in A_l]\!] \max_{j' \in B_l} \sum_{l' \in [k]} [\![j' \in B_{l'}]\!]
$$

for all $(i, j) \in [m] \times [n]$. So we have shown that there exists $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ with:

$$
c_{i,j} \leq 2 + 8 \sum_{l \in [k]} [\![i \in A_l]\!] \max_{j' \in B_l} \sum_{l' \in [k]} [\![j' \in B_{l'}]\!]
$$

for all $(i, j) \in [m] \times [n]$. By applying the same argument to $\boldsymbol{U}^\top$ instead of $\boldsymbol{U}$ we have that there exists $\boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ with:

$$
d_{i,j} \leq 2 + 8 \sum_{l \in [k]} [\![j \in B_l]\!] \max_{i' \in A_l} \sum_{l' \in [k]} [\![i' \in A_{l'}]\!]
$$

for all $(i, j) \in [m] \times [n]$, which completes the proof. ■

A.10. Proof of Theorem 10

Without loss of generality assume that n is an integer power of 2 (else just add dummy vertices to R - i.e. dummy columns to $\boldsymbol{U}$). From the tree R first construct a spine L (defined in Herbster et al. (2008)). Without loss of generality assume that the jth vertex in L is j (noting that all vertices are in $[n]$). From Herbster et al. (2008) we have:

$$\sum_{j\in[n-1]} [\![u_{i,j} \neq u_{i,j+1}]\!] \leq 2 \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!] \tag{20}$$

for all $i \in [m]$. Now construct a binary support tree B (defined in Herbster et al. (2008)) of L, letting the added vertices be the set $[2n-1] \setminus [n]$. For all $i \in [m]$ construct the vector $\tilde{\boldsymbol{u}}_i \in \mathbb{R}^{2n-1}$ such that for all $j \in [n]$ we have $\tilde{u}_{i,j} := u_{i,j}$ and for all $j \in [2n-1] \setminus [n]$ we have $\tilde{u}_{i,j} := \tilde{u}_{i,\downarrow(j)}$ where $\downarrow(j)$ is the left-child of j in the tree B. From Equation (20) we then have:

$$\sum_{(l,k)\in B} [\![\tilde{u}_{i,l} \neq \tilde{u}_{i,k}]\!] \leq \log_2(n) \sum_{j\in[n-1]} [\![u_{i,j} \neq u_{i,j+1}]\!] \leq 2\log_2(n) \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!]\,. \tag{21}$$

Let r be the maximum value of any diagonal element of the pseudoinverse of the laplacian of B. Note that it is a standard result that r is upper bounded by the resistance diameter of B, which is equal to $2\log_2(n)$.

We have from Herbster and Pontil (2006) (noting the equivalence between the kernel perceptron and the ordinary perceptron over a feature space) that there exists a matrix $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{(2n-1)\times(2n-1)}$ with:

$$\max_{j\in[2n-1]} \|\tilde{\boldsymbol{q}}_j\|^2 \leq r \tag{22}$$

such that for all vectors $\tilde{\boldsymbol{u}}$ there exists a vector $\tilde{\boldsymbol{p}}$ such that for all $j \in [2n-1]$ we have $\tilde{\boldsymbol{p}} \cdot \tilde{\boldsymbol{q}}_j = \tilde{u}_j$, and:

$$\|\tilde{\boldsymbol{p}}\|^2 = \frac{1}{r} + 4 \sum_{(l,k)\in B} [\![\tilde{u}_l \neq \tilde{u}_k]\!]\,.$$

Hence, by equation (21), there exists such a $\tilde{\boldsymbol{Q}}$, and a matrix $\tilde{\boldsymbol{P}} \in \mathbb{R}^{m\times(2n-1)}$ such that for all $i \in [m]$ we have $\tilde{\boldsymbol{p}}_i \cdot \tilde{\boldsymbol{q}}_j = u_{i,j}$ and

$$\|\tilde{\boldsymbol{p}}_i\|^2 = \frac{1}{r} + 4 \sum_{(l,k)\in B} [\![\tilde{u}_{i,l} \neq \tilde{u}_{i,k}]\!] \leq \frac{1}{r} + 8\log_2(n) \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!] \tag{23}$$

so choose such a $\tilde{\boldsymbol{Q}}$ and $\tilde{\boldsymbol{P}}$. Now define $\boldsymbol{P} \in \mathbb{R}^{m\times(2n+1)}$ and $\boldsymbol{Q} \in \mathbb{R}^{n\times(2n+1)}$ such that for all $(i,j) \in [m] \times [n]$ we have $\boldsymbol{p}_i := \tilde{\boldsymbol{p}}_i / \|\tilde{\boldsymbol{p}}_i\|$ and $\boldsymbol{p}_j := \tilde{\boldsymbol{p}}_j / \|\tilde{\boldsymbol{p}}_j\|$, noting that:

$$(\boldsymbol{p}_i \cdot \boldsymbol{q}_j) u_{i,j} = \frac{(\tilde{\boldsymbol{p}}_i \cdot \tilde{\boldsymbol{q}}_j) u_{i,j}}{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|} = \frac{u_{i,j}^2}{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|} = \frac{1}{\|\tilde{\boldsymbol{p}}_i\| \|\tilde{\boldsymbol{q}}_j\|}\,. \tag{24}$$

Then define $\boldsymbol{C} \in \mathbb{R}^{m\times n}$ such that:

$$c_{i,j} := \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2$$

for all $(i,j) \in [m]\times[n]$. By equation (24) and the fact that $\|\boldsymbol{p}_i\| = \|\boldsymbol{q}_j\| = 1$ for all $(i,j) \in [m]\times[n]$, we have that $(\boldsymbol{P},\boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ and hence $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$. By equations (22), (23) and (24) we have:

$$\begin{aligned} c_{i,j} &= \left(\frac{1}{\boldsymbol{p}_i \cdot \boldsymbol{q}_j}\right)^2 = \|\tilde{\boldsymbol{p}}_i\|^2 \|\tilde{\boldsymbol{q}}_j\|^2 \\ &\leq \|\tilde{\boldsymbol{p}}_i\|^2 r \\ &\leq \left(\frac{1}{r} + 8\log_2(n) \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!]\right) r \\ &= 1 + r8\log_2(n) \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!] \\ &= 1 + 16\log_2(n)^2 \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!] \end{aligned}$$

for all $(i,j) \in [m]\times[n]$. So we have shown that, for any value of n (not just a power of two), there exists $\boldsymbol{C} \in \mathcal{E}(\boldsymbol{U})$ such that:

$$c_{i,j} \leq 1 + 16\lceil\log_2(n)\rceil^2 \sum_{(l,k)\in R} [\![u_{i,l} \neq u_{i,k}]\!]$$

for all $(i,j) \in [m]\times[n]$. By the same argument, but with $\boldsymbol{U}^\top$ instead of $\boldsymbol{U}$, there also exists $\boldsymbol{D} \in \mathcal{E}(\boldsymbol{U})$ such that:

$$d_{i,j} \leq 1 + 16\lceil\log_2(m)\rceil^2 \sum_{(l,k)\in S} [\![u_{l,j} \neq u_{k,j}]\!]$$

for all $(i,j) \in [m]\times[n]$, as required. ■

A.11. Proof of Theorem 11

Let $L := L_\ell^\dagger$ and $\theta := \theta_\ell$. Let $S := |L|$. For all $s \in [S]$ let $t(s)$ be the s-th trial in L - i.e. such that $t(s) \in L$ and $\sum_{t\in L}[\![t \leq t(s)]\!] = s$. For all $s \in [S]$ let $\tilde{\boldsymbol{W}}_s$ be the matrix $\boldsymbol{W}_\ell^\dagger$ at the start of trial $t(s)$ and let $\tilde{\boldsymbol{X}}_s := \boldsymbol{X}_{t(s)}$. Then the values of $\{\hat{y}_{t(s)} : s \in [S]\}$ that are produced by PCMC are equal to those formed by the following algorithm:

1. $\tilde{\boldsymbol{W}}_1 \leftarrow \boldsymbol{I}$
2. For $s = 1, 2, 3 \cdots S$:
   (a) Set $\tilde{\boldsymbol{X}}_s := \frac{1}{2}(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})^\top$
   (b) Output $\hat{y}_{t(s)} := \text{sign}(\text{Tr}(\tilde{\boldsymbol{X}}_s\tilde{\boldsymbol{W}}_s) - 1)$
   (c) $\tilde{\boldsymbol{W}}_{s+1} := \exp(\ln(\tilde{\boldsymbol{W}}_s) + \frac{1}{2}\theta(y_{t(s)} - \hat{y}_{t(s)})\tilde{\boldsymbol{X}}_s)$

Let l be such that $\boldsymbol{P} \in \mathbb{R}^{m\times l}$. Let $\tilde{M} := \{s \in [S] \mid t(s) \in M(L)\}$ and let $\tilde{Z} := \{s \in [S] \mid t(s) \in Z(L,\theta,\boldsymbol{P},\boldsymbol{Q})\}$ noting that $\tilde{Z} \subseteq \tilde{M}$. Define $\tilde{\boldsymbol{P}}$ and $\tilde{\boldsymbol{Q}}$ by, for all $(i,j) \in [m]\times[n]$:

$$\tilde{\boldsymbol{p}}_i := [\,[\![i \in A(L)]\!]\boldsymbol{p}_i\,,\,[\![i \notin A(L)]\!]\boldsymbol{e}_i\,] \quad ; \quad \tilde{\boldsymbol{q}}_j = [\,[\![j \in B(L)]\!]\boldsymbol{q}_j\,,\,[\![j \notin B(L)]\!]\boldsymbol{e}_{m+j}\,]$$

where here $[\,\cdot\,,\,\cdot\,]$ concatenates two vectors. Now define $\boldsymbol{V} := [\tilde{\boldsymbol{P}}, \tilde{\boldsymbol{Q}}]$ and define $\tilde{\boldsymbol{U}} := \boldsymbol{V}\boldsymbol{V}^\top$. Given positive definite matrices $\boldsymbol{C}, \boldsymbol{D}$ of the same dimensionality we define their quantum relative entropy as:

$$\Delta(\boldsymbol{C}, \boldsymbol{D}) := \mathrm{Tr}(\boldsymbol{C}\ln(\boldsymbol{C}) - \boldsymbol{C}\ln(\boldsymbol{D}) + \boldsymbol{D} - \boldsymbol{C})\,.$$

For all $s \in [S]$ we then define:

$$\tilde{\Delta}_s := \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_s) - \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_{s+1})\,.$$

Lemma 19 *We have:*

$$\mathrm{Tr}(\tilde{\boldsymbol{U}}) = m + n$$

and:

$$\mathrm{Tr}(\tilde{\boldsymbol{U}}\ln(\tilde{\boldsymbol{U}})) \leq (|A(L)| + |B(L)|)\ln(m+n)\,.$$

Proof The first equality is true since:

$$\mathrm{Tr}(\tilde{\boldsymbol{U}}) = \sum_{i\in[m+n]} \boldsymbol{v}_i \cdot \boldsymbol{v}_i = \sum_{i\in[m+n]} \|\boldsymbol{v}_i\|^2 = \sum_{i\in[m]} \|\tilde{\boldsymbol{p}}_i\|^2 + \sum_{i\in[n]} \|\tilde{\boldsymbol{q}}_i\|^2 = m + n\,.$$

Let $\boldsymbol{\sigma} \in \mathbb{R}^{m+n}$ be the vector whose components are the eigenvalues of $\tilde{\boldsymbol{U}}$ in any order. For any $i \in [m] \setminus A(L)$ we have that $\tilde{\boldsymbol{p}}_i = [\boldsymbol{0}, \boldsymbol{e}_i]$ so by definition of $\tilde{\boldsymbol{P}}$ and $\tilde{\boldsymbol{Q}}$ we have $\tilde{\boldsymbol{p}}_i \cdot \tilde{\boldsymbol{p}}_i = 1$ and for all $i' \in [m] \setminus \{i\}$ we have $\tilde{\boldsymbol{p}}_{i'} \cdot \tilde{\boldsymbol{p}}_i = 0$ and for all $j \in [n]$ we have $\tilde{\boldsymbol{q}}_j \cdot \tilde{\boldsymbol{p}}_i = 0$. This implies that $\boldsymbol{V}\tilde{\boldsymbol{p}}_i = \boldsymbol{e}_i$ and hence that $\tilde{\boldsymbol{U}}\boldsymbol{e}_i = \boldsymbol{V}(\boldsymbol{V}^\top \boldsymbol{e}_i) = \boldsymbol{V}\tilde{\boldsymbol{p}}_i = \boldsymbol{e}_i$ so that $\boldsymbol{e}_i$ is an eigenvector of $\tilde{\boldsymbol{U}}$ with eigenvalue 1. Similarly, for all $j \in [n] \setminus B(L)$ we have that $\boldsymbol{e}_{m+j}$ is an eigenvector of $\tilde{\boldsymbol{U}}$ with eigenvalue 1. This gives us at least $(n+m) - |A(L)| - |B(L)|$ eigenvectors with eigenvalue 1. Hence, without loss of generality, assume that $\sigma_i = 1$ for all $i \in [m+n]$ with $i > |A(L)| + |B(L)|$. We have, from the first equality, that $\|\boldsymbol{\sigma}\|_1 = \mathrm{Tr}(\tilde{\boldsymbol{U}}) = m + n$ so:

$$\begin{aligned}
\sum_{i\leq|A(L)|+|B(L)|} \sigma_i &= \|\boldsymbol{\sigma}\|_1 - \sum_{i=|A(L)|+|B(L)|+1}^{m+n} \sigma_i \\
&= \|\boldsymbol{\sigma}\|_1 - (m + n - (|A(L)| + |B(L)|)) \\
&= |A(L)| + |B(L)|\,.
\end{aligned}$$

Since $\tilde{\boldsymbol{U}}$ is positive semi-definite all components of $\boldsymbol{\sigma}$ are non-negative and hence, from the first equality, we have $\ln(\sigma_i) \leq \ln(\|\boldsymbol{\sigma}\|_1) = \ln(m+n)$ for all $i \in [m+n]$. This gives us:

$$\begin{aligned}
\mathrm{Tr}(\tilde{\boldsymbol{U}}\ln(\tilde{\boldsymbol{U}})) &= \sum_{i\in[m+n]} \sigma_i \ln(\sigma_i) \\
&= \sum_{i\leq|A(L)|+|B(L)|} \sigma_i \ln(\sigma_i) + \sum_{i=|A(L)|+|B(L)|+1}^{m+n} \sigma_i \ln \sigma_i \\
&= \sum_{i\leq|A(L)|+|B(L)|} \sigma_i \ln(\sigma_i) + \sum_{i=|A(L)|+|B(L)|+1}^{m+n} \ln(1) \\
&= \sum_{i\leq|A(L)|+|B(L)|} \sigma_i \ln(\sigma_i)
\end{aligned}$$

$$\begin{aligned}&\leq \sum_{i\leq |A(L)|+|B(L)|} \sigma_i \ln(m+n)\\ &= (|A(L)|+|B(L)|)\ln(m+n)\end{aligned}$$

as required. ■

Lemma 20 *For all $s \in \tilde{M}$ we have:*

$$y_{t(s)}(\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) - 1) \geq [\![s \notin \tilde{Z}]\!]\theta\,.$$

Proof We have

$$\begin{aligned}\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) &= \frac{1}{2}\operatorname{Tr}(\boldsymbol{V}\boldsymbol{V}^\top(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})^\top)\\ &= \frac{1}{2}\operatorname{Tr}((\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})^\top\boldsymbol{V}\boldsymbol{V}^\top(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}}))\\ &= \frac{1}{2}\|\boldsymbol{V}^\top(\boldsymbol{e}_{i_{t(s)}} + \boldsymbol{e}_{m+j_{t(s)}})\|^2\\ &= \frac{1}{2}\|\boldsymbol{v}_{i_{t(s)}} + \boldsymbol{v}_{m+j_{t(s)}}\|^2\\ &= \frac{1}{2}\|\tilde{\boldsymbol{p}}_{i_{t(s)}} + \tilde{\boldsymbol{q}}_{j_{t(s)}}\|^2\,.\end{aligned}$$

Since $t(s) \in L$ we have $i_{t(s)} \in A(L)$ and hence $\tilde{\boldsymbol{p}}_{i_{t(s)}} = \boldsymbol{p}_{i_{t(s)}}$. Similarly we have $\tilde{\boldsymbol{q}}_{j_{t(s)}} = \boldsymbol{q}_{j_{t(s)}}$ so since $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ we have:

$$\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) = \frac{1}{2}\|\boldsymbol{p}_{i_{t(s)}} + \boldsymbol{q}_{j_{t(s)}}\|^2 = \frac{1}{2}\|\boldsymbol{p}_{i_{t(s)}}\|^2 + \boldsymbol{p}_{i_{t(s)}}\cdot\boldsymbol{q}_{j_{t(s)}} + \frac{1}{2}\|\boldsymbol{q}_{j_{t(s)}}\|^2 = 1 + \boldsymbol{p}_{i_{t(s)}}\cdot\boldsymbol{q}_{j_{t(s)}}$$

and hence:

$$y_{t(s)}(\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) - 1) = u_{i_{t(s)},j_{t(s)}}(\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) - 1) = u_{i_{t(s)},j_{t(s)}}\boldsymbol{p}_{i_{t(s)}}\cdot\boldsymbol{q}_{j_{t(s)}}$$

Since $(\boldsymbol{P}, \boldsymbol{Q}) \in \mathcal{N}(\boldsymbol{U})$ this is always greater than 0 and if $s \notin \tilde{Z}$ we have $t(s) \in M(L) \setminus Z(L, \theta, \boldsymbol{P}, \boldsymbol{Q})$ so $|\boldsymbol{p}_{i_{t(s)}}\cdot\boldsymbol{q}_{j_{t(s)}}| \geq \theta$ as required. ■

Lemma 21 *For all $s \in \tilde{M}$ we have:*

$$\tilde{\Delta}_s \geq \theta y_{t(s)}\operatorname{Tr}(\tilde{\boldsymbol{U}}\tilde{\boldsymbol{X}}_s) + (1 - e^{\theta y_{t(s)}})\operatorname{Tr}(\tilde{\boldsymbol{W}}_s\tilde{\boldsymbol{X}}_s)\,.$$

Proof As in Herbster et al. (2016) Lemma A.3. ■

Lemma 22 *For all $s \in [S]$:*

- *If $s \notin \tilde{M}$ then $\tilde{\Delta}_s = 0$.*
- *If $s \in \tilde{Z}$ then $\tilde{\Delta}_s \geq (c-1)\theta^2$.*

- *If* $s \in \tilde{M} \setminus \tilde{Z}$ *then* $\tilde{\Delta}_s \geq c\theta^2$.

Proof As in Herbster et al. (2016) Lemma A.4, multiplying by $m+n$ throughout. Note that this proof utilises lemmas 20 and 21 above. ■

By Lemma 19 we have:

$$\begin{aligned}
\Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_1) &= \Delta(\tilde{\boldsymbol{U}}, \boldsymbol{I}) \\
&= \mathrm{Tr}(\tilde{\boldsymbol{U}} \ln(\tilde{\boldsymbol{U}})) - \mathrm{Tr}(\tilde{\boldsymbol{U}} \ln(\boldsymbol{I})) + \mathrm{Tr}(\boldsymbol{I}) - \mathrm{Tr}(\tilde{\boldsymbol{U}}) \\
&\leq (|A(L)| + |B(L)|) \ln(m+n) - \mathrm{Tr}(\tilde{\boldsymbol{U}} \ln(\boldsymbol{I})) + \mathrm{Tr}(\boldsymbol{I}) - (m+n) \\
&= (|A(L)| + |B(L)|) \ln(m+n) - \mathrm{Tr}(\tilde{\boldsymbol{U}}\boldsymbol{0}) + (m+n) - (m+n) \\
&= (|A(L)| + |B(L)|) \ln(m+n)
\end{aligned}$$

so since $\Delta(\boldsymbol{C}, \boldsymbol{D}) \geq 0$ for all positive definite matrices $\boldsymbol{C}, \boldsymbol{D}$ we have, by Lemma 22 that:

$$\begin{aligned}
(|A(L)| + |B(L)|) \ln(m+n) &\geq \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_1) \\
&\geq \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_1) - \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_{S+1}) \\
&= \sum_{s \in [S]} (\Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_s) - \Delta(\tilde{\boldsymbol{U}}, \tilde{\boldsymbol{W}}_{s+1})) \\
&= \sum_{s \in [S]} \tilde{\Delta}_s \\
&\geq |\tilde{Z}|(c-1)\theta^2 + |\tilde{M} \setminus \tilde{Z}| c\theta^2
\end{aligned}$$

so:

$$|\tilde{M} \setminus \tilde{Z}| \leq \frac{\ln(m+n)}{c\theta^2}(|A(L)| + |B(L)|) + \frac{1-c}{c}|\tilde{Z}|$$

and hence:

$$\begin{aligned}
|M(L)| &= |\tilde{M}| \\
&= |\tilde{M} \setminus \tilde{Z}| + |\tilde{Z}| \\
&\leq \frac{\ln(m+n)}{c\theta^2}(|A(L)| + |B(L)|) + \frac{1}{c}|\tilde{Z}| \\
&= \frac{\ln(m+n)}{c\theta^2}(|A(L)| + |B(L)|) + \frac{1}{c}|Z(L, \theta, \boldsymbol{P}, \boldsymbol{Q})|
\end{aligned}$$

as required. ■