Predicting the Labelling of a Graph via Minimum p-Seminorm Interpolation

Mark Herbster **Guy Lever**
Department of Computer Science
University College London
Gower Street, London WC1E 6BT, England, UK
{*m.herbster, g.lever*}*@cs.ucl.ac.uk*

Abstract

We study the problem of predicting the labelling of a graph. The graph is given and a trial sequence of (vertex,label) pairs is then incrementally revealed to the learner. On each trial a vertex is queried and the learner predicts a boolean label. The true label is then returned. The learner's goal is to minimise mistaken predictions. We propose *minimum p-seminorm interpolation* to solve this problem. To this end we give a p-seminorm on the space of graph labellings. Thus on every trial we predict using the labelling which *minimises* the p-seminorm and is also *consistent* with the revealed (vertex, label) pairs. When $p = 2$ this is the *harmonic energy minimisation* procedure of [22], also called (Laplacian) *interpolated regularisation* in [1]. In the limit as $p \to 1$ this is equivalent to predicting with a label-consistent mincut. We give mistake bounds relative to a label-consistent mincut and a resistive cover of the graph. We say an edge is *cut* with respect to a labelling if the connected vertices have disagreeing labels. We find that minimising the p-seminorm with $p = 1 + \epsilon$ where $\epsilon \to 0$ as the graph diameter $D \to \infty$ gives a bound of $\mathcal{O}(\Phi^2 \log D)$ versus a bound of $\mathcal{O}(\Phi D)$ when $p = 2$ where Φ is the number of cut edges.

1 Introduction

We study the problem of predicting the labelling of a graph in the online learning framework. Consider the following game for predicting the labelling of a graph: *Nature* presents a graph; *nature* queries a vertex v_{i_1}; the *learner* predicts the label of the vertex $\hat{y}_1 \in \{-1, 1\}$; *nature* presents a label y_1; *nature* queries a vertex v_{i_2}; the *learner* predicts $\hat{y}_2$; and so forth. The learner's goal is to minimise the total number of mistakes $M = |\{t : \hat{y}_t \neq y_t\}|$. If nature is adversarial, the learner will always mispredict, but if nature is regular or simple, there is hope that a learner may make only a few mispredictions. Thus, a central goal of on-line learning is to design algorithms whose total mispredictions can be bounded relative to the complexity of nature's labelling.

In previous work [11, 9] we used a norm induced by the graph Laplacian to predict the labelling of a graph with algorithms such as the Perceptron. In [15] it was shown that the perceptron, "online SVM"s and similar algorithms applied to the problem of learning *sparse* linear classifiers, suffer from the limitation that there exist example sequences such that these algorithms incur mistakes linearly in the dimension of the examples. These lower bounds should be contrasted to upper bounds of algorithms such as Winnow [18] and the p-norm Perceptron [8, 7] which are only logarithmic in the dimension of the examples. An analogous observation for the graph labelling problem [10] demonstrated that there exists an n-vertex graph with a single cut edge for which Laplacian 2-seminorm interpolation incurs $\theta(\sqrt{n})$ mistakes.

Inspired by the results for the p-norm perceptron's ability to learn sparse concepts in $\mathbb{R}^n$, we consider a similar idea for building classifiers on graphs. We thus introduce a family of seminorms defined on the vertices of a graph – we term them Laplacian p-seminorms which include the smoothness functional of [1, 22] and the label-consistent graph cut [2] as limiting cases. We present an online algorithm for learning concepts defined on graphs based upon minimum p-seminorm interpolation. We derive a mistake bound for this algorithm in which the graph cut of a labelling is the measure of the complexity of the learning task. In the graph setting the dual seminorm gives rise to a generalisation of the notion of resistance between graph vertices [16, 6], which we term p-resistance. The p-resistance is a natural measure of similarity between graph vertices and it features as the "structural" term in our mistake bound. We give a brief survey of its fundamental properties by extending a well-known analogy with resistive networks.

We demonstrate that, in natural cases, the optimal choice for the parameter p results in an algorithm which lies between the mincut ($p = 1$) of [2] and the method of minimising the smoothness functional ($p = 2$) of [1, 22]. In a further parallel with the behaviour of the p-norm Perceptron we demonstrate that we can choose the parameter p (using only information available a-priori to the learner) to ensure a performance guarantee which is logarithmic with regard to graph diameter. The bound also decreases with the edge connectivity of the graph or clusters thereof as a consequence of the p-resistance term.

2 Background and preliminaries

If $\boldsymbol{z} \in \mathbb{R}^n$ then let $\|\boldsymbol{z}\|_p := \sqrt[p]{\sum_{i=1}^n |z_i|^p}$ denote the p-norm when $p \in [1, \infty)$. More generally, if $\boldsymbol{\Psi} : \mathbb{R}^n \to \mathbb{R}^m$ is any linear map we define the associated $(\boldsymbol{\Psi}, p)$-seminorm

as $\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p} := \|\boldsymbol{\Psi u}\|_p$. If $\{\boldsymbol{0}\} = \{\boldsymbol{u} \in \mathbb{R}^n : \boldsymbol{\Psi u} = \boldsymbol{0}\}$ then $\|\cdot\|_{\boldsymbol{\Psi},p}$ defines a norm since we have a unique minimal vector. Given a seminorm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ the *dual* seminorm $\|\cdot\|^* : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is defined on the vector space of linear functionals $Z : \mathbb{R}^n \to \mathbb{R}$ as

$$\|Z\|^* := \sup_{\boldsymbol{w} \in \mathbb{R}^n} \left\{ \frac{|Z(\boldsymbol{w})|}{\|\boldsymbol{w}\|} \right\} = [\inf_{\boldsymbol{w} \in \mathbb{R}^n} \{\|\boldsymbol{w}\| : Z(\boldsymbol{w}) = 1\}]^{-1}. \quad (1)$$

We immediately recover the useful inequality

$$|Z(\boldsymbol{w})| \leq \|Z\|^* \|\boldsymbol{w}\| .$$

The canonical basis vectors of $\mathbb{R}^n$ we denote as $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ with corresponding functionals $E_i(\boldsymbol{w}) := \boldsymbol{e}_i^\top \boldsymbol{w}$.

Given a set $X \subseteq \mathcal{X}$, a *cover* of X is a collection $\mathcal{C} = \{X_i\}_{i=1}^k$ of subsets $X_i \subseteq \mathcal{X}$ such that $X \subseteq \cup_{i=1}^k X_i$. For a given symmetric *discrepancy* function $d : X \times X \to \mathbb{R}$ ($d(x,y) = d(y,x)$) and any $\rho > 0$, the *covering number* $\mathcal{N}(X, \rho, d(\cdot,\cdot))$ of X is the cardinality of the smallest cover $\mathcal{C}$ such that for each $X_i \in \mathcal{C}$ we have $d(x,x') \leq \rho$ if $x, x' \in X_i$.

A graph $\mathcal{G} = (V, E)$ is a collection of vertices $V = \{v_1, \ldots, v_n\}$ joined by connecting (possibly weighted) edges. Denote $i \sim j$ whenever v_i and v_j are connected by an edge. We consider *undirected* graphs so that $E := \{(i,j) | i \sim j\}$ is the set of unordered pairs of adjacent vertex indexes. Associated with each edge $(i,j) \in E$ is a weight $A_{ij} > 0$ and $A_{ij} = 0$ if $(i,j) \notin E$, so that $\boldsymbol{A}$ is the (weighted) symmetric *adjacency matrix*. We say that $\mathcal{G}$ is *unweighted* if $\boldsymbol{A} \in \{0,1\}^{n \times n}$.

We say $\mathcal{G}'$ is a *subgraph* of $\mathcal{G}$ whenever $V'_{\mathcal{G}} \subseteq V_{\mathcal{G}}$ and $E'_{\mathcal{G}} \subseteq E_{\mathcal{G}}$ and we write $\mathcal{G}' \subseteq \mathcal{G}$. If $V'_{\mathcal{G}} \subseteq V_{\mathcal{G}}$ then the *induced subgraph* is $(V'_{\mathcal{G}}, E'_{\mathcal{G}})$ with $E'_{\mathcal{G}} := \{(i,j) \in E_{\mathcal{G}} : v_i, v_j \in V'_{\mathcal{G}}\}$.

A *path graph* $\mathcal{P}$ is a graph of the form $V_{\mathcal{P}} = \{v_0, v_1 ... v_n\}$, $E_{\mathcal{P}} = \{(0,1), (1,2) ... (n-1, n)\}$ and we define the length, $\ell(\mathcal{P})$, of any path $\mathcal{P}$ by $\ell(\mathcal{P}) = \sum_{(i,j) \in E_{\mathcal{P}}} \frac{1}{A_{ij}}$. The *distance* between any two vertices $v_i, v_j \in V_{\mathcal{G}}$ is the length of the shortest path containing v_i and v_j,

$$\delta(i,j) = \min_{\{\mathcal{P} \subseteq \mathcal{G} : v_i, v_j \in V_{\mathcal{P}}\}} \ell(\mathcal{P})$$

and is equal to ∞ if no path exists. We define the *diameter* of $\mathcal{G}$, $D(\mathcal{G}) = \max_{i,j} \delta(i,j)$. In this paper, we generally consider *connected* graphs (that is, graphs in which a path connects any two vertices).

A *labelling* $\boldsymbol{u} \in \mathbb{R}^n$ of an n-vertex graph $\mathcal{G}$ is viewed as a function $\boldsymbol{u} : V_{\mathcal{G}} \to \mathbb{R}$ defined on the vertices of $\mathcal{G}$ whereby u_i corresponds to the label of v_i. If $\mathcal{G} = (V, E = \{(i_1, j_1), \ldots, (i_m, j_m)\})$ is a graph then an associated *edge map* (weighted oriented incidence matrix) $\boldsymbol{\Psi}_{\mathcal{G}} : \mathbb{R}^n \to \mathbb{R}^m$ (with p implicit) is a linear map such that

$$\boldsymbol{\Psi}_{\mathcal{G}} \boldsymbol{u} = (A_{i_1 j_1}^{\frac{1}{p}} (u_{i_1} - u_{j_1}), \ldots, A_{i_m j_m}^{\frac{1}{p}} (u_{i_m} - u_{j_m}))^\top . \quad (2)$$

When $p = 2$, the $n \times n$ matrix $\mathbf{G} = \boldsymbol{\Psi}_{\mathcal{G}}^\top \boldsymbol{\Psi}_{\mathcal{G}}$ is the well-known graph *Laplacian*. We introduce a class of *Laplacian p-seminorms* defined on the space of graph labellings: if $\boldsymbol{u} \in \mathbb{R}^n$ then

$$\|\boldsymbol{u}\|_{\mathcal{G},p} := \|\boldsymbol{u}\|_{\boldsymbol{\Psi}_{\mathcal{G}},p} = \left(\sum_{(i,j) \in E_{\mathcal{G}}} A_{ij} |u_i - u_j|^p \right)^{\frac{1}{p}} . \quad (3)$$

These seminorms generalise the commonly used "smoothness functional" $\boldsymbol{u}^T \boldsymbol{G} \boldsymbol{u}$ [1, 22] and as such measure the complexity of graph labellings. When the labelling is restricted to $\boldsymbol{u} \in \{-1, 1\}^n$ we say that edge (i,j) is *cut* if $u_i \neq u_j$ and we define the *weighted cut size* of $\boldsymbol{u}$ as

$$\Phi_{\mathcal{G}}(\boldsymbol{u}) := \frac{1}{2^p} \|\boldsymbol{u}\|_{\mathcal{G},p}^p = \frac{1}{2^p} \sum_{(i,j) \in E} A_{ij} |u_i - u_j|^p . \quad (4)$$

The cut-size is independent of p and if the graph is unweighted it is just the number of cut edges.

We will use the dual norm $\|\cdot\|_{\mathcal{G},p}^*$ to give a discrepancy $r_{\mathcal{G},p}(\cdot,\cdot)$ between vertices by identifying vertices v_i and v_j with the functionals E_i and E_j so that

$$r_{\mathcal{G},p}(i,j) = \left(\|E_i - E_j\|_{\mathcal{G},p}^* \right)^p .$$

When $p = 2$ there is an established natural connection [6] between graphs and resistive networks where each edge $(i,j) \in E_{\mathcal{G}}$ is viewed as a resistor with resistance $\frac{1}{A_{ij}}$. The *effective resistance* $r_{\mathcal{G}}(i,j) = r_{\mathcal{G},2}(i,j)$ is the potential difference needed to induce a unit current flow between v_i and v_j. The *p-resistance* (*diameter*) of a graph $\mathcal{G}$ is defined $R_p(\mathcal{G}) := \max_{\{v_i, v_j \in V_{\mathcal{G}}\}} r_{\mathcal{G},p}(i,j)$ ($R(\mathcal{G}) = R_2(\mathcal{G})$). In this paper the notion of (*effective) p-resistance* will be a key to our bounds and is further developed in Section 4.1.

2.1 Previous work

The problem of learning a labeling of a graph is a natural problem in the online learning setting, as well as a foundational technique for a variety of semi-supervised learning methods [2, 17, 22, 1]. One practical application of graph labelling is found in the image segmentation problem. Here for example we are given an image and we distinguish the foreground from the background. The user may select a set of vertices ("pixels") and the system should then return a segmentation, i.e. a classification of every pixel into foreground or background. The graph's topology corresponds to pixel connectivity and the edges are weighted according to interpixel similarity (e.g., color). Such a system based on p-seminorm interpolation was considered in [21]. For another example, in the online setting, consider a system which serves advertisements on web pages. The web pages may be identified with the vertices of a graph and the edges as links between pages. The online prediction problem is then that, at a given time t the system may receive a request to serve an advertisement on a particular web page. For simplicity, we assume that there are two alternatives to be served: either advertisement "A" or advertisement "B". The system then interprets the feedback as the label and then may use this information in responding to the next request to predict an advertisement for a requested web page.

The problem of predicting the labelling of a graph in the online framework was first considered in [12] and a mistake bound for the kernel perceptron was given in [11, Theorem 4.2 (with $b = R(\mathcal{G}); c = 0$)] of

$$|\mathcal{M}| \leq 8 \Phi_{\mathcal{G}}(\boldsymbol{u}) R(\mathcal{G}) + 2,$$

where $\boldsymbol{u}$ is any labelling consistent with the trial sequence.

In [9] the Pounce on-line prediction technique was developed to exploit any cluster structure in a graph. The algorithm achieves the mistake bound

$$|\mathcal{M}| \leq \mathcal{N}(X, \rho, \sqrt{r_{\mathcal{G}}}) + 4 \Phi_{\mathcal{G}}(\boldsymbol{u}) \rho^2 + 1,$$

for any $\rho > 0$. Here, $\boldsymbol{u} \in \mathbb{R}^n$ is any labelling consistent with the trial sequence, $X = \{v_{i_1}, v_{i_2}, \dots v_{i_T}\} \subseteq V$ is the set of inputs and the covering number $\mathcal{N}(X, \rho, \sqrt{r_{\mathcal{G}}})$ is the minimum number of vertex sets of resistance diameter no greater than ρ^2 required to cover X (see Section 2). The Pounce algorithm therefore captures the notion of cluster structure through a graph cover of low resistance vertex sets.

In [10] a limitation of existing methods for predicting the labelling of a graph (including an online version of the common and well-motivated method of minimising the smoothness functional given by (3) when $p = 2$) was identified; n-vertex graph constructions exist for which the algorithms incur (at least) $\theta(\sqrt{\Phi_{\mathcal{G}}(\boldsymbol{u})n})$ mistakes. It was demonstrated that any unweighted graph can be embedded into a path graph in such a way that an efficient Bayes optimal classifier used to predict the labelling of the embedding (and, therefore, of the underlying graph) obtains a mistake bound which grows only logarithmically in the size of the graph

$$|\mathcal{M}| \leq 2\Phi_{\mathcal{G}}(\boldsymbol{u}) \max\left[0, \log_2\left(\frac{n-1}{2\Phi_{\mathcal{G}}(\boldsymbol{u})}\right)\right] + \frac{2\Phi_{\mathcal{G}}(\boldsymbol{u})}{\ln 2} + 1. \tag{5}$$

This algorithm, however, involves the corruption of the graph structure resulting in a drawback: the method does not exploit graph connectivity – in fact the mistake bound (5) improves if the graph is replaced by any spanning tree – and is therefore not demonstrably suitable for the case of dense or clustered data. A further algorithm to utilise an embedding of $\mathcal{G}$ into a simpler structure was presented in [4] and here the reduction is to a tree $\mathcal{T}$. A mistake bound of

$$|\mathcal{M}| \leq \mathcal{O}(\Phi_{\mathcal{T}}(\boldsymbol{u}) \log D(\mathcal{C}))$$

is derived, where here $\Phi_{\mathcal{T}}(\boldsymbol{u})$ is the cut size of the true labelling $\boldsymbol{u}$ on $\mathcal{T}$ and $D(\mathcal{C})$ is the maximum diameter of any cluster (unitarily labelled) of vertices which $\mathcal{T}$ is partitioned into by $\boldsymbol{u}$.

A goal of research in this area is to present an algorithm which fully exploits cluster structure and connectivity in graphs *and* obtains a logarithmic performance guarantee. In this paper we present an algorithm with a mistake bound in terms of a revealing resistance feature and demonstrate that this is upper bounded by a logarithmic function of the graph diameter. The algorithm therefore exploits cluster structure and connectivity but is also suitable in the case in which a graph exhibits a sparse structure or large diameter.

3 Minimum (Ψ, p)-seminorm interpolation

Given the problem of predicting a labelling of a set of objects, a natural approach is to specify a norm on the labelling of those objects and to choose a labelling which is then both consistent and minimal in norm; this approach is known as *minimum norm interpolation*. Recalling Section 2, in this paper we investigate interpolation with $(\boldsymbol{\Psi}, p)$-*seminorms*, $\|\cdot\|_{\boldsymbol{\Psi},p}$, which are specified by choosing a linear map $\boldsymbol{\Psi} : \mathbb{R}^n \to \mathbb{R}^m$ and a $p \in (1, 2]$. In the case when $p = 2$ and when $\boldsymbol{\Psi}$ has a rank of n this is equivalent to using using the Euclidean norm induced by the kernel matrix $K = (\boldsymbol{\Psi}^\top \boldsymbol{\Psi})^{-1}$. The intention is that $\boldsymbol{\Psi}$ is chosen so that the $(\boldsymbol{\Psi}, p)$-seminorm captures the geometry of the problem in question, and in our application it will capture the geometry of a graph.

Given a $(\boldsymbol{\Psi}, p)$-seminorm and a sequence of online trials $t \in \{1, 2, 3, ...\}$ in which (index,label) pairs (i_t, y_t) are revealed, our algorithm (see Figure 1) maintains a weight vector $\boldsymbol{w}_t \in \mathbb{R}^n$ such that $\operatorname{sgn}(\boldsymbol{e}_{i_t}^\top \boldsymbol{w}_t)$ is the hypothesised label for indexed object i_t at trial t. On trial t, the weight vector is updated by choosing that vector consistent with all previous examples[1] which attains the least $(\boldsymbol{\Psi}, p)$-seminorm, if there are multiple minimisers an arbitrary vector is chosen[2].

Parameters: A linear map $\boldsymbol{\Psi} : \mathbb{R}^n \to \mathbb{R}^m$ and $p \in (1, 2]$
Initialization: $\boldsymbol{w}_1 = \mathbf{0}$; $\mathcal{M} = \{\}$
Input: $\{(i_t, y_t)\}_{t=1}^{\ell} \in (\mathbb{N}_n \times \{-1, 1\})^{\ell}$
for $t = 1, \dots, \ell$ **do**
  Receive: $i_t \in \{1, \dots, n\}$
  Predict: $\hat{y}_t = \operatorname{sign}(\boldsymbol{e}_{i_t}^\top \boldsymbol{w}_t)$
  Receive: y_t
  if $\hat{y}_t \neq y_t$ **then** $\mathcal{M} = \mathcal{M} \cup \{t\}$
  $\boldsymbol{w}_{t+1} = \operatorname{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \{\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p} : u_{i_1} = y_1, \dots, u_{i_t} = y_t\}$
end

Figure 1: Minimum $(\boldsymbol{\Psi}, p)$-seminorm interpolation

In the case when $\boldsymbol{\Psi}$ is an edge map, the indices naturally correspond to the vertices on a graph. We bound the mistakes of our interpolation algorithm in the following theorem.

Theorem 1 *The number of mistakes, $|\mathcal{M}|$, incurred by minimum $(\boldsymbol{\Psi}, p)$-seminorm interpolation, for any $\rho > 0$, is bounded by*

$$|\mathcal{M}| \leq \mathcal{N}(X, \rho, d_{\boldsymbol{\Psi},p}) + \frac{\rho^2 \|\boldsymbol{u}\|_{\boldsymbol{\Psi},p}^2}{p-1} \tag{6}$$

where $\boldsymbol{u} \in \mathbb{R}^n$ is any labelling such that $u_{i_t} = y_t \; \forall t \leq \ell$, and $\mathcal{N}(X, \rho, d_{\boldsymbol{\Psi},p})$ is the covering number of the input set $X = \{i_1, i_2, ..., i_\ell\}$ relative to the distance

$$d_{\boldsymbol{\Psi},p}(i, j) := \|E_i - E_j\|_{\boldsymbol{\Psi},p}^*. \tag{7}$$

The bound above is for the general case, a proof is given in Appendix A. In the following we will study the case corresponding to prediction of the labelling of a graph where $\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p}^2$ will correspond to a function of the cut size (see (4)) of the labelling $\boldsymbol{u}$ and $d_{\boldsymbol{\Psi},p}(i, j)$ will be identified with a measure closely related to resistance in an electrical network.

4 Interpolation on a graph

We proceed to our intended application of predicting the labelling of a given graph $\mathcal{G}$ by choosing $\boldsymbol{\Psi}$ to be an edge map $\boldsymbol{\Psi}_{\mathcal{G}}$ of $\mathcal{G}$ (recall (2)). If we denote the adjacency of $\mathcal{G}$ by $\boldsymbol{A}$, $(\boldsymbol{\Psi}_{\mathcal{G}}, p)$-seminorm interpolation on $\mathcal{G}$ is therefore the process of choosing the labelling $\boldsymbol{u}$ of $\mathcal{G}$ which minimises the seminorm (recalling (3))

$$\|\boldsymbol{u}\|_{\boldsymbol{\Psi}_{\mathcal{G}},p} = \left(\sum_{(i,j) \in E_{\mathcal{G}}} A_{ij} |u_i - u_j|^p\right)^{\frac{1}{p}}$$

[1] The conservative version of the algorithm where a vector is chosen consistent with only the "mistaken" examples obtains the same bound as Theorem 1.

[2] If $\boldsymbol{\Psi}$ is the edge map of a connected graph this will never occur.

subject to the constraints imposed by the revealed vertex labels. The dual norm term (7) of our mistake bound for $(\mathbf{\Psi}, p)$-seminorm interpolation now corresponds to the following generalization of effective resistance.

Definition 2 *Given a graph $\mathcal{G}$, we define the (effective) p-resistance between any two vertices $v_a, v_b \in V_\mathcal{G}$ as*

$$r_{\mathcal{G},p}(a,b) := (||E_a - E_b||^*_{\mathcal{G},p})^p. \tag{8}$$

Thus when $p = 2$ this is the usual effective resistance and as $p \to 1$ then $r_{\mathcal{G},p}(s,t) \to \frac{1}{\text{“st-mincut”}}$. We will see that for $1 < p \leq 2$ effective p-resistance provides a natural measure of similarity between vertices on a graph.

Rewriting Theorem 1 with the substitution (8) we now have the following corollary, which is our main result.

Corollary 3 *After ℓ trials we have, for any $\rho > 0$,*

$$|\mathcal{M}| \leq \mathcal{N}(X, \rho, r_{\mathcal{G},p}) + \frac{\rho^{\frac{2}{p}} \|\boldsymbol{u}\|^2_{\mathcal{G},p}}{p-1} \tag{9}$$

where $\boldsymbol{u} \in \mathbb{R}^n$ is any labelling of $\mathcal{G}$ such that $u_{i_t} = y_t \; \forall t \leq \ell$, $p \in (1,2]$, and $\mathcal{N}(X, \rho, r_{\mathcal{G},p})$ is the covering number of the input set $X = \{v_{i_1}, v_{i_2}, ...v_{i_\ell}\}$ relative to the p-resistance $r_{\mathcal{G},p}$.

We proceed to develop an interpretation of the bound (9) to culminate in Corollary 10. The norm of the classifier $\|\boldsymbol{u}\|^2_{\mathcal{G},p}$ is relatively simple to interpret while the properties of the p-resistance measure $r_{\mathcal{G},p}$ are less immediate. We therefore now establish an instructive theory of the p-resistance which will both clarify the bound above and provide guidance on the tuning of the parameter p.

4.1 Theory of p-resistive networks

We now build on a popular connection between the graph labelling problem and the problem of identifying the potential at the nodes of an electric network derived from the graph [22, 6]. We describe the notion of a network as parallel to a partially labelled graph, in which each edge is a resistive conduit along which electric charge flows between vertices. The label u_i of a vertex v_i is equivalent to it's electric *potential* (or voltage). A partial labelling constrains the potential on the corresponding subset of vertices in the network, through which current then flows along edges according to the laws of the electric network theory. The foundation of our theory here differs from standard theory in a single respect – energy is produced in resistors according to a purely hypothetical formulation of power. This results in changes to other familiar key concepts, such as Ohm's law.

A *p-resistive network* $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$ consists of an n-vertex weighted connected graph $\mathcal{G} = (V, E)$ with adjacency $\boldsymbol{A}$, a set $\mathcal{S} = \{(v_{i_1}, y_1), \ldots, (v_{i_\ell}, y_\ell)\} \in (V_\mathcal{G} \times \mathbb{R})^\ell$ of $0 \leq \ell \leq n$ feasible potential constraints and a constant $p \in (1, 2]$. The potential constraints can be viewed as (the effect of) voltage sources applied to the relevant vertices. Denote by $V_\mathcal{S}$ the set of constrained vertices. The resistance of an edge, $\pi_{ij} := \frac{1}{A_{ij}} \in (0, \infty)$, measures the resistance of (i,j) to current flow and is constant. Given a network $\mathcal{C}$ a *state* is an assignment of potentials $\boldsymbol{u} \in \mathbb{R}^n$ to $V_\mathcal{G}$. In the following we will additionally define for any network, a *power* $P(\mathcal{C}, \cdot) : \mathbb{R}^n \to [0, \infty)$, a *current* $\boldsymbol{I}(\mathcal{C}) : V \times V \to \mathbb{R}$ satisfying $I_{ij} = -I_{ji}$, and $I_{ij} = 0$ whenever $A_{ij} = 0$, and when $\mathcal{G}$ is clear from the context we will abbreviate the effective p-resistance $r_{\mathcal{G},p}$ to r_p.

4.1.1 Fundamental properties

To draw a parallel with our graph labelling problem we define the *power* of potential state $\boldsymbol{u}$ as

$$P(\boldsymbol{u}) := \sum_{(i,j)\in E} \frac{|u_i - u_j|^p}{\pi_{ij}}. \tag{10}$$

and the corresponding power of any edge (i,j) as

$$P_{ij}(\boldsymbol{u}) := \frac{|u_i - u_j|^p}{\pi_{ij}}.$$

The standard electric network theory corresponds to the choice $p = 2$, and all other choices result in hypothetical theories. Determining the labelling with minimal p-seminorm (3) subject to certain boundary constraints is equivalent to determining the potential state which minimises (10) under the same boundary constraints. Given a network $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$, if the potential constraints $\mathcal{S} = \{(v_{i_1}, y_1), \ldots, (v_{i_\ell}, y_\ell)\} \neq \emptyset$ then let $\boldsymbol{w}(\mathcal{C})$ denote the unique minimiser

$$\boldsymbol{w}(\mathcal{C}) = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \{P(\boldsymbol{u}) : u_{i_1} = y_1, \ldots, u_{i_\ell} = y_\ell\}.$$

A p-resistive network operates according to the principal of minimising (10) and so a set of potential constraints $\mathcal{S}$ induces the minimal potential state $\boldsymbol{w}(\mathcal{C})$ on the network. The *power of a network* $\mathcal{C}$ is therefore defined as the power of the minimal feasible state

$$P(\mathcal{C}) := \min_{\boldsymbol{u}\in\mathbb{R}^n} \{P(\boldsymbol{u}) : u_{i_1} = y_1, \ldots, u_{i_\ell} = y_\ell\}.$$

At the minimum we have

$$\begin{aligned} \frac{\partial P(\boldsymbol{u})}{\partial u_i}\Big|_{\boldsymbol{u}=\boldsymbol{w}} &= 0 \quad v_i \notin V_\mathcal{S} \\ \sum_{j:j\sim i} \frac{|w_i - w_j|^{p-1}\text{sgn}(w_i - w_j)}{\pi_{ij}} &= 0 \quad v_i \notin V_\mathcal{S}. \end{aligned} \tag{11}$$

We define the *current* from vertex v_i to v_j of a network

$$I_{ij}(\mathcal{C}) := \frac{|w_i - w_j|^{p-1}\text{sgn}(w_i - w_j)}{\pi_{ij}} \tag{12}$$

(if $p = 2$ this is *Ohm's law*) and the *net current* from vertex v_i as

$$I_i := \sum_{j:j\sim i} I_{ij}\,.$$

Since $\pi_{ij} \geq 0$ we see that current flows from vertices with high potential to those with low potential. We see that (11) is *Kirchoff's current law* for $\boldsymbol{I}$

$$0 = I_i \quad v_i \notin V_\mathcal{S} \tag{13}$$

and that we can alternatively express power via *Joule's law*

$$P_{ij}(\boldsymbol{w}) = (w_i - w_j)I_{ij}. \tag{14}$$

Lemma 4 *Given a network $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$ with potential constraints $\mathcal{S} = \{(v_a, y_a), (v_b, y_b)\}$, then*

$$P(\mathcal{C}) = (w_a - w_b)I_a \tag{15}$$

where $\boldsymbol{w}$ and $\boldsymbol{I}$ are the minimal potential state and the current induced by $\mathcal{S}$.

Proof: The power of a network is sum of the power along the edges $P(\mathcal{C}) = \sum_{(i,j)\in E} P_{ij}(\boldsymbol{w})$ and thus by Joule's law (14) we have

$$\begin{aligned} P(\mathcal{C}) &= \sum_{(i,j)\in E} (w_i - w_j)I_{ij} \\ &= \sum_i \sum_{j:j<i} w_i I_{ij} - \sum_j \sum_{i:i>j} w_j I_{ij} \\ &= \sum_i \sum_{j:j<i} w_i I_{ij} + \sum_i \sum_{j:j>i} w_i I_{ij} \\ &= \sum_j w_a I_{aj} + \sum_j w_b I_{bj} + \sum_{i:i\neq a,b} \sum_j w_i I_{ij} \end{aligned}$$

and the result follows since $I_a = \sum_j I_{aj} = -\sum_j I_{bj}$ and $\sum_j I_{ij} = 0 \; \forall i \neq a, b$. ■

We now demonstrate that the construction (8) can indeed naturally be interpreted as a resistance feature in our electric network analogy, via an identity similar to Ohm's Law relating potential, current and effective p-resistance.

Lemma 5 *Given a network $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$ with potential constraints $\mathcal{S} = \{(v_a, y_a), (v_b, y_b)\}$ and $y_a \neq y_b$, then*

$$P(\mathcal{C}) = \frac{|w_a - w_b|^p}{r_p(a,b)}, \tag{16}$$

and

$$r_p(a,b) = \frac{|w_a - w_b|^{p-1}\mathrm{sgn}(w_a - w_b)}{I_a}, \tag{17}$$

where $\boldsymbol{w}$ and $\boldsymbol{I}$ are the minimal potential state and the current induced by $\mathcal{S}$.

Proof: We have, by the definition of power:

$$\begin{aligned} P(\mathcal{C}) &= \min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \|\boldsymbol{u}\|_{\mathcal{G},p}^p : u_a = y_a, u_b = y_b \right\} \\ &= |y_a - y_b|^p \min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \|\boldsymbol{u}\|_{\mathcal{G},p}^p : u_a - u_b = 1 \right\}. \end{aligned}$$

Substituting (1) into (8) gives

$$r_p(a,b) = \left(\min_{\boldsymbol{u}\in\mathbb{R}^n} \left\{ \|\boldsymbol{u}\|_{\mathcal{G},p}^p : u_a - u_b = 1 \right\} \right)^{-1},$$

now equation (16) follows since $w_a = y_a$ and $w_b = y_b$. Finally if we apply Lemma 4 by substituting $P(\mathcal{C}) = (w_a - w_b)I_a$ into (16) we obtain (17). ■

4.1.2 Bounding the p-resistance

Blackbox principles in electric circuit theory are useful tools that allow the simplification of complex networks. In the p-resistive framework we give analogues of the classic "series" (Lemma 6) and "parallel" laws (Lemma 7). The fact that we can chain together sequential applications of these laws is guaranteed by the seemingly intuitive Thevenin-type theorem (Theorem 8).

Lemma 6 *(Resistors in series) Consider a path graph $\mathcal{P}$, with $V_\mathcal{P} = \{v_1, v_2 ... v_n\}, E_\mathcal{P} = \{(1,2), (2,3) ... (n-1, n)\}$ and edge resistance π_{ij} for each $i \sim j$. Then*

$$r_p(1,n) = \left(\sum_{i=1}^{n-1} \pi_{i,i+1}^{\frac{1}{p-1}} \right)^{p-1}.$$

Proof: Given a network $\mathcal{C} = (\mathcal{P}, \mathcal{S}, p)$ with potential constraints $\mathcal{S} = \{(v_1, y_1), (v_n, y_n)\}$ let $\boldsymbol{w}$ and $\boldsymbol{I}$ denote the minimal potential state and current induced on $\mathcal{C}$. We have, from Lemma 5 and (12)

$$\begin{aligned} w_1 - w_n &= \sum_{i=1}^{n-1} w_i - w_{i+1} \\ |I_1|^{\frac{1}{p-1}} r_p(1,n)^{\frac{1}{p-1}} &= \sum_{i=1}^{n-1} |I_{i,i+1}|^{\frac{1}{p-1}} \pi_{i,i+1}^{\frac{1}{p-1}} \end{aligned}$$

and the result follows since, by (13), we have that $I_1 = I_{i,i+1}$ for $i < n$. ■

Lemma 7 *(Resistors in parallel) Consider a multigraph $\mathcal{G}$ with two vertices $V_\mathcal{G} = \{v_a, v_b\}$ joined by m resistive edges with resistances $\{\pi_k\}_{k=1}^m$. Then*

$$r_p(a,b) = \left(\sum_{k=1}^m \frac{1}{\pi_k} \right)^{-1}$$

Proof: Given a network $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$ with potential constraints $\mathcal{S} = \{(v_a, y_a), (v_b, y_b)\}$ let $\boldsymbol{w}$ denote the minimal potential state on $\mathcal{C}$. Then by (16) we have the following identity for the power $P(\mathcal{C})$

$$\frac{|w_a - w_b|^p}{r_p(a,b)} = \sum_{k=1}^m \frac{|w_a - w_b|^p}{\pi_k}$$

and the result follows immediately. ■

We first define the notion of a resistive *unit* $\mathcal{U} = (V_\mathcal{U}, E_\mathcal{U})$ as any combination of resistors and vertices with two *terminal* vertices $V_\mathcal{U}^T = \{v_a, v_b\} \subseteq V_\mathcal{U}$. We refer to the non-terminal vertices $V_\mathcal{U}^I = V_\mathcal{U} \backslash V_\mathcal{U}^T$ as the *interior* vertices. Any unit $\mathcal{U}$ can be treated as a component in a larger graph $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$, such that $\mathcal{U} \subseteq \mathcal{G}$ and whenever $v \in V_\mathcal{U}, v' \in V_\mathcal{G} \backslash V_\mathcal{U}$ and $v \sim v'$ then $v \in V_\mathcal{U}^T$.

Theorem 8 *(Thevenin) Any resistive unit $\mathcal{U}$ with two terminals v_a and v_b and with effective p-resistance $r_{\mathcal{U},p}(a,b)$ is electrically identical to a single edge with p-resistance $\pi_{ab} = r_{\mathcal{U},p}(a,b)$. In particular, in any given network in which $\mathcal{U}$ is a component and $V_\mathcal{U}^I$ is unconstrained we can "black box" $\mathcal{U}$, and replace it with a single edge of p-resistance $r_{\mathcal{U},p}(a,b)$ without affecting current or potential in the external network.*

Proof: Consider a network $\mathcal{C} = (\mathcal{G}, \mathcal{S}, p)$ in which $\mathcal{U}$ is a component of an n-vertex graph $\mathcal{G} = (V_\mathcal{G}, E_\mathcal{G})$ with adjacency $\boldsymbol{A}$. Suppose that the non-empty potential constraints $\mathcal{S}$ are defined on a subset of vertices $V_\mathcal{S} \subseteq V_\mathcal{G} \backslash V_\mathcal{U}^I$ not in the interior of $\mathcal{U}$. Denote by $\boldsymbol{w}$ and $\boldsymbol{I}$ the minimal feasible potential state and current, and by $P(\mathcal{C})$ the induced power. Define the power produced across $\mathcal{U}$ by potential state $\boldsymbol{u} \in \mathbb{R}^n$ as $P_\mathcal{U}(\boldsymbol{u}) = \sum_{(i,j)\in E_\mathcal{U}} A_{ij}|u_i - u_j|^p$.

Consider a second network $\mathcal{C}' = (\mathcal{G}', \mathcal{S}, p)$ formed by replacing $\mathcal{U}$ with a single edge (a, b); $V_{\mathcal{G}'} = V_\mathcal{G} \backslash V_\mathcal{U}^I$, $E_{\mathcal{G}'} = (E_\mathcal{G} \backslash E_\mathcal{U}) \cup \{(a, b)\}$. Let $\pi_{ab} = r_{\mathcal{U},p}(a, b)$, $|V_{\mathcal{G}'}| = n'$ and denote the adjacency of $\mathcal{G}'$ by $\boldsymbol{A}'$. Let $\boldsymbol{w}'$ denote the minimal feasible potential state induced by $\mathcal{S}$ on $\mathcal{C}'$.

The potential at no vertex $v \in V_\mathcal{U}^I$ is constrained by $\mathcal{S}$ and so $P_\mathcal{U}(\boldsymbol{w})$ is equal to the power produced across $\mathcal{U}$ when it is considered as an isolated circuit with the terminal vertices constrained to $\{(v_a, w_a), (v_b, w_b)\}$. Since such a circuit satisfies the conditions for Lemma 5 we have

$$\begin{aligned} P_\mathcal{U}(\boldsymbol{w}) &= \frac{|w_a - w_b|^p}{r_{\mathcal{U},p}(a, b)} \\ &= \frac{|w_a - w_b|^p}{\pi_{ab}}. \end{aligned}$$

Thus $P_\mathcal{U}(\boldsymbol{w})$ is always identical to the power produced across a single edge with resistance $\pi_{ab} = r_{\mathcal{U},p}(a, b)$ and

$$\begin{aligned} P(\mathcal{C}') &= \min_{\boldsymbol{u}\in\mathbb{R}^{n'}} \{ \sum_{(i,j)\in E_{\mathcal{G}'}} |u_i - u_j|^p A'_{ij} : \mathcal{S}\} \\ &= \min_{\boldsymbol{u}\in\mathbb{R}^{n'}} \{ \sum_{(i,j)\in E_\mathcal{G}\backslash E_\mathcal{U}} |u_i - u_j|^p A_{ij} + \frac{|u_a - u_b|^p}{\pi_{ab}} : \mathcal{S}\} \\ &= \min_{\boldsymbol{u}\in\mathbb{R}^{n'}} \{ \sum_{(i,j)\in E_\mathcal{G}\backslash E_\mathcal{U}} |u_i - u_j|^p A_{ij} + \frac{|u_a - u_b|^p}{r_{\mathcal{U},p}(a, b)} : \mathcal{S}\} \\ &= \min_{\boldsymbol{u}\in\mathbb{R}^{n}} \{ \sum_{(i,j)\in E_\mathcal{G}} |u_i - u_j|^p A_{ij} : \mathcal{S}\} \\ &= P(\mathcal{C}) \end{aligned} \tag{18}$$

It is then sufficent to notice that $\boldsymbol{w}'$ must be identical to $\boldsymbol{w}$ on $V_\mathcal{G} \backslash V_\mathcal{U}^I$ since by (18) they then produce the same (minimal) power: $P_\mathcal{C}(\boldsymbol{w}) = P_{\mathcal{C}'}(\boldsymbol{w}')$. That current on the external circuits is identical follows from (12).

■

We demonstrate that the effective p-resistance satisfies an equivalent of Rayleigh's monotonicity law – suppose that the weighting of some edge of $\mathcal{G}$ is increased (equivalently, its resistance is decreased) or a new edge created, then the effective p-resistance between any two vertices of $\mathcal{G}$ does not increase.

Lemma 9 *(Rayleigh's Monotonicity Principal) Given $\mathcal{G}$ with adjacency matrix $\boldsymbol{A}$. Let $\mathcal{G}'$, with adjacency $\boldsymbol{A}'$, be identical to $\mathcal{G}$ except for the increase in the weight of one arbitrary edge (a, b), so that $A'_{ab} = A'_{ba} = A_{ab} + \delta$ for $\delta > 0$. Then for arbitrary vertices i and j,*

$$r_{\mathcal{G},p}(i, j) \geq r_{\mathcal{G}',p}(i, j)\,.$$

Proof: Given any $v_i, v_j \in V_\mathcal{G}$, let

$$\boldsymbol{w} = \operatorname*{argmin}_{\boldsymbol{u}\in\mathbb{R}^n} \{\|\boldsymbol{u}\|^p_{\mathcal{G},p} : u_i - u_j = 1\}.$$

Suppose that we can find a labelling $\boldsymbol{w}'$ of $\mathcal{G}'$ such that $w'_i - w'_j = 1$ and $\|\boldsymbol{w}'\|^p_{\mathcal{G}',p} < \|\boldsymbol{w}\|^p_{\mathcal{G},p}$, then note that

$$\begin{aligned} &\sum_{(k,\ell)\in E_\mathcal{G}} |w'_k - w'_\ell|^p A_{k\ell} \\ &\qquad = \sum_{(k,\ell)\in E_{\mathcal{G}'}} |w'_k - w'_\ell|^p A'_{k\ell} - |w'_a - w'_b|^p \delta \\ &\qquad \leq \sum_{(k,\ell)\in E_{\mathcal{G}'}} |w'_k - w'_\ell|^p A'_{k\ell} \\ &\qquad < \sum_{(k,\ell)\in E_\mathcal{G}} |w_k - w_\ell|^p A_{k\ell} \end{aligned}$$

which contradicts the minimality of $\boldsymbol{w}$. Hence

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} \{\|\boldsymbol{u}\|^p_{\mathcal{G}',p} : u_i - u_j = 1\} \geq \|\boldsymbol{w}\|^p_{\mathcal{G},p}$$

from which (8) implies

$$r_{\mathcal{G}',p}(i, j) \leq r_{\mathcal{G},p}(i, j).$$

■

Further we also have monotonicity in "p" so that for a graph $\mathcal{G}$ and vertices i and j if $p \leq s$ then

$$r_{\mathcal{G},p}(i, j) \leq r_{\mathcal{G},s}(i, j)\,.$$

4.2 Analysing the mistake bound for unweighted graphs

We are now better equipped with an understanding of effective p-resistance to analyse the mistake bound, Corollary 3. We see, through Lemmas 6 and 7, that p-resistance is a discrepancy measure which captures both connectivity and distance. Since it is difficult to evaluate the behaviour of (9) through p-resistance directly, we choose a more tractable approximation: we generalize the notion of graph diameter to that of (unweighted) wide diameter [13]. This approximation captures connectivity in the graph structure.

The k-wide distance $\delta_k(i, j)$ is the minimum value ℓ such that there exists k edge disjoint paths each containing v_i and v_j of length no more than ℓ (and $\delta_k(i, j) = \infty$ if no such k paths exist). We then define the k-wide diameter $\Delta_k(\mathcal{G}) := \max_{i,j}(\delta_k(i, j))$. Thus $\Delta_1(\mathcal{G})$ is just the usual diameter and if

$$\Phi^0_\mathcal{G} := \min_{\boldsymbol{u}\in\{-1,1\}^n} \{\Phi_\mathcal{G}(\boldsymbol{u}) : \Phi_\mathcal{G}(\boldsymbol{u}) \geq 1\}$$

then by Menger's theorem [5] then there exists $\Phi^0_\mathcal{G}$ edge-disjoint paths between all pairs of vertices. Thus if $k \leq \Phi^0_\mathcal{G}$ then $\Delta_k(\mathcal{G}) \leq n$. We can now bound the p-resistance diameter of an unweighted graph $\mathcal{G}$ by

$$R_p(\mathcal{G}) \leq \frac{\Delta_k(\mathcal{G})^{p-1}}{k}. \tag{19}$$

This follows immediately from application of resistors in parallel and series laws (Lemmas 6 and 7) to the set of k edge disjoints paths determined by the wide diameter $\Delta_k(\mathcal{G})$ and an application of Rayleigh's monotonicity principle (Lemma 9). We observe that (19) becomes tight as $p \to 1$ hence,

$$\lim_{p\to 1} R_p(\mathcal{G}) = \frac{1}{\Phi^0_\mathcal{G}}\,.$$

In the following we use the upper bound (19) to investigate the mistake bound (9). In [10] it was demonstrated that the case $p = 2$ (which is an online version of the harmonic energy minimisation of [22, 1]) suffers a limitation – there exist graphs for which the algorithm makes $\theta(\sqrt{|V_{\mathcal{G}}|})$ mistakes. It has been demonstrated that simple online algorithms with a logarithmic mistake bound exist [10, 4]. In Section 4.2.2 we will demonstrate that it is possible to choose p to ensure that $(\mathbf{\Psi}_{\mathcal{G}}, p)$-seminorm interpolation achieves a logarithmic guarantee.

4.2.1 The choice of p

A natural question arises: how does the behaviour of the $(\mathbf{\Psi}_{\mathcal{G}}, p)$-seminorm interpolation algorithm differ for various choices of p? To begin an investigation into this question we first deduce a mistake bound for the unweighted graph case in terms of a graph's wide diameter, and consider a simple tuning of p for unweighted graphs. For any vertex set partition $V_1 \cup \ldots \cup V_N = V_{\mathcal{G}}$ with induced subgraphs $G_1, \ldots, G_N$ of maximum wide diameter $\Delta_k := \max\{\Delta_k(G_i) : i = 1, \ldots, N\}$ we have as an immediate consequence of Corollary 3

$$|\mathcal{M}| \leq N + \frac{4\Delta_k^2}{p-1}\left(\frac{\Phi(\boldsymbol{u})}{k\Delta_k}\right)^{\frac{2}{p}}, \tag{20}$$

for any $\boldsymbol{u} \in \{-1, 1\}^n$ correct on all trials. For the purpose of investigating the dependence of the bound (20) on the parameter p, we consider the hypothetical situation in which the graph cut $\Phi(\boldsymbol{u})$ is known to the learner a-priori and consider tuning (20) with regard to p. Note that, for $k\Delta_k > e^2\Phi(\boldsymbol{u})$ the quantity $\frac{1}{p-1}\left(\frac{\Phi(\boldsymbol{u})}{k\Delta_k}\right)^{\frac{2}{p}}$ is minimised when

$$p = p^* = \log\left(\frac{k\Delta_k}{\Phi(\boldsymbol{u})}\right) - \sqrt{\left(\log\left(\frac{k\Delta_k}{\Phi(\boldsymbol{u})}\right)\right)^2 - 2\log\left(\frac{k\Delta_k}{\Phi(\boldsymbol{u})}\right)}$$

and we have that $1 < p^* < 2$. Of course, the value of $k\Delta_k$ is dependent upon the (optimal) choice of graph partition. Very generally, when the diameter of a graph is large relative to the cut, lower values of p optimise (20). The situation is not simple, however, due to the connectivity element; below we demonstrate a dense, clustered graph for which a small choice of p is equally reasonable.

4.2.2 A simple tuning

We now give a simpler tuning (near-optimal) which will be used to evaluate the behaviour of p-seminorm interpolation in instructive cases. In a parallel with the logarithmic behaviour of the p-norm Perceptron, we show that it is possible to choose p (using information known to the learner a-priori) to ensure a performance guarantee which is logarithmic in the graph diameter.

Corollary 10 *Given the task of predicting the labelling of any unweighted, connected graph $\mathcal{G} = (V, E)$ in the online framework, the number of mistakes, $|\mathcal{M}|$, incurred by minimum $(\mathbf{\Psi}_{\mathcal{G}}, p)$-seminorm interpolation with $p := \frac{c}{c-1}$ is bounded by*

$$|\mathcal{M}| \leq \begin{cases} N + \frac{4e^2\Phi^2(\boldsymbol{u})[\log(k\Delta_k) - \log(\widehat{\Phi}) - 1]}{k^2} & \frac{k\Delta_k}{\widehat{\Phi}} > e^2 \\ N + \frac{4\Phi(\boldsymbol{u})\Delta_k}{k} & \frac{k\Delta_k}{\widehat{\Phi}} \leq e^2 \end{cases}$$

where $c = \min(\log[\frac{k\Delta_k}{\widehat{\Phi}}], 2)$ and $V_1 \cup \ldots \cup V_N = V_{\mathcal{G}}$ is any vertex set partition with induced subgraphs $G_1, \ldots, G_N$ of maximum wide diameter $\Delta_k := \max\{\Delta_k(G_i) : i = 1, \ldots, N\}$, $\widehat{\Phi}$ is any constant $1 \leq \widehat{\Phi} \leq \Phi(\boldsymbol{u})$ and $\boldsymbol{u} \in \{-1, 1\}^n$ is any labelling consistent with the trial sequence.

Note immediately that by choosing $k = 1, \widehat{\Phi} = 1$, for $\Delta_1 = \max_i D(\mathcal{G}_i) > e^2$, we recover a mistake bound which is a logarithmic function of the graph diameter. In the following we consider three examples with varying degrees of connectivity. The *tree* , a prototypically sparse graph, is minimally connected with $k = 1$. The $2m$-vertex dense *barbell*, an idealized model of two clusters, has connectivity $k = m - 1$. Finally the mD-vertex *cylinder* has an intermediate connectivity $k = m$. This intermediate case more generally includes graphs with spatially extended clusters whose internal connectivity equals or exceeds the cut between clusters. The bounds for these intermediately connected graphs uniformly improve on the results in [9, 10, 4].

Tree graph

Consider a tree. We take $N = 1$, $k = 1, \Delta_k = D = \max_i D(\mathcal{G}_i)$ in Corollary 10. For $\frac{D}{\widehat{\Phi}} > e^2$ the first tuning ($p < 2$) in Corollary 10 is preferred and we derive

$$|\mathcal{M}| \leq 1 + 4e^2\Phi^2(\boldsymbol{u})[\log(D) - \log(\widehat{\Phi}) - 1].$$

For $\frac{D}{\widehat{\Phi}} \leq e^2$ we derive, from the second tuning ($p = 2$)

$$|\mathcal{M}| \leq 1 + 4\Phi(\boldsymbol{u})D.$$

Barbell graph

Consider the barbell graph: two m-cliques joined by Φ connecting cut edges. We take $N = 2, \Delta_k = 2, k = m - 1$ in Corollary 10. For $\frac{2(m-1)}{\widehat{\Phi}} > e^2$ the first tuning ($p < 2$) in Corollary 10 is preferred and we derive

$$|\mathcal{M}| \leq 2 + \frac{4e^2\Phi^2(\boldsymbol{u})[\log(2(m-1)) - \log(\widehat{\Phi}) - 1]}{(m-1)^2}.$$

For $\frac{2(m-1)}{\widehat{\Phi}} \leq e^2$ we derive, from the second tuning ($p = 2$)

$$|\mathcal{M}| \leq 2 + \frac{8\Phi(\boldsymbol{u})}{m-1}.$$

Note that a bound of 2 is optimal for this barbell graph labelling problem.

Cylinder graph

Consider the "cylindrical" graph that is the cartesian product of an m-clique with a path graph of D vertices. This cylinder may be visualized as D "aligned" cliques. We assume the cylinder is labeled with two classes by an m-edge cut that partitions into two cylinders. Assuming $D > e^2 - 1$ then choosing $p = 1 + \frac{1}{\log(D+1)-1}$ (with $N = 1$, $k = m$, and $\Delta_k = D + 1$) and substituting into Corollary 10 we derive

$$|\mathcal{M}| \leq 4e^2 \log(D + 1).$$

If instead we tune with $p = 2$ we have $|\mathcal{M}| \leq 5 + 4D$. Further this bound improves on the "spine" method in [10] which has a bound of $O(k \log D)$ for this problem.

5 Discussion

We have presented an algorithm for predicting the labelling of a graph which achieves bounds of a similar form to those of the p-norm perceptron [8]. As with the p-norm perceptron there is a direct argument that gives bounds which scale logarithmically with the dimension (n) of the input space. We refined these "$\mathcal{O}(\log n)$" to "$\mathcal{O}(\log D)$" bounds in section 4.2 using the geometrical results on p-resistive networks from 4.1. The bounds may be further improved by recognizing that the diameter D of the input space is replaceable by the diameter of the balls that constitute a "cover" of the inputs. This was accomplished by adapting the methods of [9] to a p-norm framework. We note the following open problem. As discussed for trees we obtain a bounds of $\mathcal{O}(\Phi^2 \log D)$. In [10] and in [4] efficient online algorithms were proposed with mistake bounds of $\mathcal{O}(\Phi \log \frac{n}{\Phi} + \Phi)$ and $\mathcal{O}(\Phi \log D)$ respectively. The drawbacks of these algorithms are that they are not able to fully exploit additional connectivity in non-tree graphs as typified by barbell or cylinder graphs. This leaves as an open problem the discovery of an algorithm that can obtain $\mathcal{O}(\Phi \log D)$ on trees but also exploit edge-connectivity as typified by Corollary 10.

Acknowledgments

We would like to thank the anonymous referees for useful comments. We also thank Vladimir Kolmogorov for useful discussions pointing out the reference [21] to us and Andreas Argyriou for useful discussions and comments on the manuscript. Finally, we would like to thank the PASCAL 2 European network of excellence for supporting this work.

References

[1] M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *Proc. of the 17–th Annual Conf. on Learning Theory (COLT'04)*, Banff, Alberta, 2004.

[2] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proc. 18th International Conf. on Machine Learning*, pages 19–26. Morgan Kaufmann, San Francisco, CA, 2001.

[3] L. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Physics*, 7:200–217, 1967.

[4] N. Cesa-Bianchi, C. Gentile, and F. Vitale. Fast and optimal prediction on a labeled tree. COLT 2009 (to appear), 2009.

[5] R. Diestel. *Graph Theory*, volume 173 of *Graduate Texts in Mathematics*. Springer-Verlag, Heidelberg, third edition, 2005.

[6] P. G. Doyle and J. L. Snell. Random walks and electric networks, 2000.

[7] C. Gentile. The robustness of the p-norm algorithms. *Machine Learning*, 53(3):265–299, 2003.

[8] A. J. Grove, N. Littlestone, and D. Schuurmans. General convergence results for linear discriminant updates. In *Proc. 10th Annu. Conf. on Comput. Learning Theory*, pages 171–183. ACM, 1997.

[9] M. Herbster. Exploiting cluster-structure to predict the labeling of a graph. In *The 19th International Conference on Algorithmic Learning Theory*, pages 54–69, 2008.

[10] M. Herbster, G. Lever, and M. Pontil. Online prediction on large diameter graphs. In *NIPS*, 2008.

[11] M. Herbster and M. Pontil. Prediction on a graph with a perceptron. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 577–584. MIT Press, Cambridge, MA, 2007.

[12] M. Herbster, M. Pontil, and L. Wainer. Online learning over graphs. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 305–312, New York, NY, USA, 2005. ACM Press.

[13] D. Hsu. On container width and length in graphs, groups, and networks. *IEICE Trans. Fundamental of Electronics, Comm., and Computer Sciences*, A(4):668–680, 1994.

[14] S. M. Kakade, S. Shalev-Schwartz, and A. Tewari. On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization. Unpublished Manuscript, `http://ttic.uchicago.edu/~shai/papers/KakadeShalevTewari09.pdf`, 2009.

[15] J. Kivinen, M. K. Warmuth, and P. Auer. The perceptron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. *Artificial Intelligence*, 97:325–343, December 1997.

[16] D. Klein and M. Randić. Resistance distance. *Journal of Mathematical Chemistry*, 12(1):81–95, 1993.

[17] R. I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. In *ICML 2002*, 2002.

[18] N. Littlestone. Learning when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2:285–318, 1988.

[19] R. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[20] S. Shalev-Shwartz. *Online Learning: Theory, Algorithms, and Applications*. PhD thesis, The Hebrew University of Jerusalem, 2002.

[21] D. Singaraju, L. Grady, and R. Vidal. P-brush: Continuous valued mrfs with normed pairwise distributions for image segmentation. In *CVPR 2009*.

[22] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *20-th International Conference on Machine Learning (ICML-2003)*, pages 912–919, 2003.

A Mistake bound analysis (Theorem 1)

We introduce the Bregman divergence in Section A.1 then in A.2 we show that the minimum p-seminorm interpolation algorithm is equivalent to successive projections with regard to a Bregman divergence and we complete our proof in A.3.

A.1 Bregman divergence

Bregman [3] introduced the *Bregman divergence* for convex programming.

Definition 11 *Let $F : \mathbb{R}^n \to \mathbb{R}$ be a $\mathscr{C}^2$ convex function. Denote by $D_F(\boldsymbol{u}, \boldsymbol{v})$ the Bregman divergence w.r.t. F;*

$$D_F(\boldsymbol{u}, \boldsymbol{v}) = F(\boldsymbol{u}) - F(\boldsymbol{w}) - (\boldsymbol{u} - \boldsymbol{w}) \cdot \boldsymbol{\nabla} F(\boldsymbol{w}). \quad (21)$$

The Bregman divergence is generally defined in terms of a *strictly* convex potential function F where "strictness" ensures the uniqueness of a projection. In our application we will use the nonstrictly convex potential $F(\boldsymbol{v}) = \|\boldsymbol{v}\|^2_{\boldsymbol{\Psi},p}$ and thus projection (see (22)) will not necessarily be unique. The Bregman divergence is nonnegative as the convexity of F guarantees that the first order approximation $F(\boldsymbol{u}) \approx F(\boldsymbol{w}) + (\boldsymbol{u}-\boldsymbol{w}) \cdot \boldsymbol{\nabla} F(\boldsymbol{w})$ is not an overestimate. We will use the following notation $D_p := D_{\|\cdot\|_p^2}$ and $D_{\boldsymbol{\Psi},p} := D_{\|\cdot\|^2_{\boldsymbol{\Psi},p}}$.

We define the projection of $\boldsymbol{w}$ onto a non-empty set $\mathcal{U} \subseteq \mathbb{R}^n$ with respect to D_F as

$$\mathcal{P}_F(\mathcal{U}; \boldsymbol{w}) := \operatorname*{argmin}_{\boldsymbol{u} \in \mathcal{U}} D_F(\boldsymbol{u}, \boldsymbol{w}) \,. \tag{22}$$

We note that the argmin is not necessarily unique.

Lemma 12 *If $\mathcal{U} \subseteq \mathbb{R}^n$ is a nonempty affine set and $\boldsymbol{w} \in \mathbb{R}^n$, then $\mathcal{P}_{\boldsymbol{\Psi},p}(\mathcal{U}; \boldsymbol{w})$ is non-empty.*

Proof: We recall that a *direction of recession* of a convex function is any direction in which the function is non-increasing [19, p. 69]. We observe that any direction of recession $\boldsymbol{x}$ of $D_{\boldsymbol{\Psi},p}(\cdot, \boldsymbol{w})$ is exactly one such that $\boldsymbol{\Psi}\boldsymbol{x} = 0$ and in these directions $D_{\boldsymbol{\Psi},p}(\cdot, \boldsymbol{w})$ is constant. It then follows that $\mathcal{P}_{\boldsymbol{\Psi},p}(\mathcal{U}; \boldsymbol{w})$ is non-empty by [19, Theorem 27.3] which in particular guarantees that a continuous convex function on $\mathbb{R}^n$ attains its minima on a given affine constraint set if the function is constant in every common direction of recession between the function and the constraint set. ■

The following is the well-known pythagorean equality for Bregman divergences.

Lemma 13 *If $\boldsymbol{w}' \in \mathbb{R}^n$ is a projection of $\boldsymbol{w} \in \boldsymbol{R}^n$ to the affine set $\mathcal{U} \subseteq \mathbb{R}^n$ with regard to the Bregman divergence D_F, then $\forall \boldsymbol{u} \in \mathcal{U}$ we have*

$$D_F(\boldsymbol{u}, \boldsymbol{w}) = D_F(\boldsymbol{w}', \boldsymbol{w}) + D_F(\boldsymbol{u}, \boldsymbol{w}'). \tag{23}$$

Proof: Let $\mathcal{U} = \cap_{i=1}^k \{\boldsymbol{u} : \boldsymbol{u} \cdot \boldsymbol{x}_i = y_i\}$. By expanding D_F in (23) we obtain the equivalent form

$$(\boldsymbol{\nabla} F(\boldsymbol{w}) - \boldsymbol{\nabla} F(\boldsymbol{w}')) \cdot (\boldsymbol{u} - \boldsymbol{w}') = 0. \tag{24}$$

Recalling the method of Lagrange multipliers to compute $\boldsymbol{w}'$, we note that the unconstrained minimum of the Lagrangian

$$L(\boldsymbol{\lambda}, \boldsymbol{v}) = D_F(\boldsymbol{v}, \boldsymbol{w}) + \sum_{i=1}^k \lambda_i (\boldsymbol{x}_i \cdot \boldsymbol{v} - y_i)$$

occurs at $\boldsymbol{v} = \boldsymbol{w}'$. Thus

$$\begin{aligned} \mathbf{0} &= \boldsymbol{\nabla}_{\boldsymbol{v}} L(\boldsymbol{\lambda}, \boldsymbol{v}) \mid_{\boldsymbol{v}=\boldsymbol{w}'} \\ &= \boldsymbol{\nabla} F(\boldsymbol{w}') - \boldsymbol{\nabla} F(\boldsymbol{w}) + \sum_{i=1}^k \lambda_i \boldsymbol{x}_i \end{aligned}$$

Thus

$$\begin{aligned} (\boldsymbol{\nabla} F(\boldsymbol{w}) - \boldsymbol{\nabla} F(\boldsymbol{w}')) \cdot (\boldsymbol{u} - \boldsymbol{w}') &= (\sum_{i=1}^k \lambda_i \boldsymbol{x}_i) \cdot (\boldsymbol{u} - \boldsymbol{w}') \\ &= 0 \end{aligned}$$

as required. ■

We build on the following lemma, which requires the linearity of $\boldsymbol{\Psi}$, to prove the important Lemma 15.

Lemma 14 *Given a linear map $\boldsymbol{\Psi}$ then*

$$D_{\boldsymbol{\Psi},p}(\boldsymbol{u}, \boldsymbol{w}) = D_p(\boldsymbol{\Psi}\boldsymbol{u}, \boldsymbol{\Psi}\boldsymbol{w}).$$

Proof: As $\|\boldsymbol{z}\|_{\boldsymbol{\Psi},p} = \|\boldsymbol{\Psi}\boldsymbol{z}\|_p$ we have, by applying the chain rule,

$$\begin{aligned} D_{\boldsymbol{\Psi},p}(\boldsymbol{u}, \boldsymbol{w}) &= \|\boldsymbol{u}\|^2_{\boldsymbol{\Psi},p} - \|\boldsymbol{w}\|^2_{\boldsymbol{\Psi},p} - (\boldsymbol{u}-\boldsymbol{w}) \cdot \boldsymbol{\nabla}_{\boldsymbol{z}} \, \|\boldsymbol{z}\|^2_{\boldsymbol{\Psi},p} \Big|_{\boldsymbol{z}=\boldsymbol{w}} \\ &= \|\boldsymbol{\Psi}\boldsymbol{u}\|_p^2 - \|\boldsymbol{\Psi}\boldsymbol{w}\|_p^2 - (\boldsymbol{u}-\boldsymbol{w}) \cdot \boldsymbol{\nabla}_{\boldsymbol{z}} \, \|\boldsymbol{\Psi}\boldsymbol{z}\|_p^2 \Big|_{\boldsymbol{z}=\boldsymbol{w}} \\ &= \|\boldsymbol{\Psi}\boldsymbol{u}\|_p^2 - \|\boldsymbol{\Psi}\boldsymbol{w}\|_p^2 - \boldsymbol{\Psi}(\boldsymbol{u}-\boldsymbol{w}) \cdot \boldsymbol{\nabla}_{\boldsymbol{z}'} \, \|\boldsymbol{z}'\|_p^2 \Big|_{\boldsymbol{z}'=\boldsymbol{\Psi}\boldsymbol{w}} \\ &= D_p(\boldsymbol{\Psi}\boldsymbol{u}, \boldsymbol{\Psi}\boldsymbol{w}) \end{aligned}$$

■

The following lemma is inspired directly by arguments upper bounding the quadratic remainder term in the Taylor's series expansion of the squared p-norm in [8]. We will need only the first inequality.

Lemma 15

$$(p-1)\|\boldsymbol{w}' - \boldsymbol{w}\|^2_{\boldsymbol{\Psi},p} \leq D_{\boldsymbol{\Psi},p}(\boldsymbol{w}', \boldsymbol{w}) \quad p \in (1, 2] \tag{25}$$

$$D_{\boldsymbol{\Psi},p}(\boldsymbol{w}', \boldsymbol{w}) \leq (p-1)\|\boldsymbol{w}' - \boldsymbol{w}\|^2_{\boldsymbol{\Psi},p} \quad p \in [2, \infty) \tag{26}$$

Proof: We first recall the Hölder inequality. If $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n$ and $\frac{1}{r} + \frac{1}{s} = 1$, then

$$\sum_{i=1}^n |a_i b_i| \leq \|\boldsymbol{a}\|_r \|\boldsymbol{b}\|_s \qquad r \in (1, \infty) \tag{27}$$

Now, if $\boldsymbol{\xi} = \boldsymbol{w}' - \boldsymbol{w}$ then, for $p \geq 2$ by Taylor's theorem there is some point $\boldsymbol{\zeta} \in \mathbb{R}^n$ such that:

$$||\boldsymbol{w}'||_p^2 - ||\boldsymbol{w}||_p^2 - \nabla \|\boldsymbol{z}\|_p^2 \mid_{\boldsymbol{z}=\boldsymbol{w}} \cdot \boldsymbol{\xi} = \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 \|\boldsymbol{z}\|_p^2}{\partial z_i \partial z_j} \right|_{\boldsymbol{z}=\boldsymbol{\zeta}} \xi_i \xi_j$$

$$D_p(\boldsymbol{w}', \boldsymbol{w}) = \frac{1}{2} \sum_{ij} \left. \frac{\partial^2 (\|\boldsymbol{z}\|_p^2)}{\partial z_i \partial z_j} \right|_{\boldsymbol{z}=\boldsymbol{\zeta}} \xi_i \xi_j$$

We have

$$\frac{\partial (||\boldsymbol{z}||_p^2)}{\partial z_i} = 2||\boldsymbol{z}||_p^{2-p} z_i^{p-1} \operatorname{sgn}(z_i) \,,$$

and for $i \neq j$,

$$\begin{aligned} \frac{\partial^2 (||\boldsymbol{z}||_p^2)}{\partial z_i \partial z_j} &= \frac{\partial}{\partial z_j} \left(2||\boldsymbol{z}||_p^{2-p} z_i^{p-1} \operatorname{sgn}(z_i) \right) \\ &= 2(2-p)||\boldsymbol{z}||_p^{2-2p} (z_i z_j)^{p-1} \operatorname{sgn}(z_i z_j) \,, \end{aligned}$$

and,

$$\frac{\partial^2 (||\boldsymbol{z}||_p^2)}{\partial z_i^2} = 2(2-p)\|\boldsymbol{z}\|_p^{2-2p} |z_i|^{2p-2} + 2(p-1)\|\boldsymbol{z}\|_p^{2-p} |z_i|^{p-2} \,.$$

Thus,

$$D_p(\boldsymbol{w}', \boldsymbol{w}) = (2-p)||\boldsymbol{\zeta}||_p^{2-2p} \sum_{i,j=1}^n \xi_i \xi_j (\zeta_i \zeta_j)^{p-1} \mathrm{sgn}(z_i z_j) + (p-1)\|\boldsymbol{\zeta}\|_p^{2-p} \sum_{i=1}^n \xi_i^2 |\zeta_i|^{p-2}$$
$$= (2-p)||\boldsymbol{\zeta}||_p^{2-2p} \left[\sum_{i=1}^n \xi_i \zeta_i^{p-1}\right]^2 + (p-1)\|\boldsymbol{\zeta}\|_p^{2-p} \sum_{i=1}^n \xi_i^2 |\zeta_i|^{p-2}.$$

For $p \geq 2$ the first term here is not positive while the second term is bounded above with equation (27) with $r = \frac{p}{2}, s = \frac{p}{p-2}$ giving,

$$D_p(\boldsymbol{w}', \boldsymbol{w}) \leq (p-1)\|\boldsymbol{\xi}\|_p^2 \qquad p \geq 2. \tag{28}$$

This is equivalent to the $(p-1)$-strong smoothness of the function $\frac{1}{2}||\cdot||_p^2$ with respect to the norm $||\cdot||_p$ (for a discussion of strong smoothness and strong convexity see [14, 20]). This function has Fenchel conjugate $\frac{1}{2}||\cdot||_q^2$, where $\frac{1}{p} + \frac{1}{q} = 1$, and by the duality of strong convexity and strong smoothness [14] we therefore have that $\frac{1}{2}||\cdot||_q^2$ is $(q-1)$-strongly convex w.r.t. $||\cdot||_q$, and so

$$D_p(\boldsymbol{w}', \boldsymbol{w}) \geq (p-1)||\boldsymbol{\xi}||_p^2 \qquad 1 < p \leq 2. \tag{29}$$

Finally, since $\|\boldsymbol{z}\|_{\boldsymbol{\Psi},p} = \|\boldsymbol{\Psi}\boldsymbol{z}\|_p$ an application of Lemma 14 to (28) and (29) gives the result. ■

A.2 Successive Bregman projection and interpolation

We prove that minimum $(\boldsymbol{\Psi}, p)$-seminorm interpolation is equivalent to the sequential composition of Bregman projections in Corollary 17. First we show that Bregman projections to affine sets compose using the following well-known lemma.

Lemma 16 *If U_1 and U_2 are affine sets and $U_2 \subseteq U_1$ then*

$$\mathcal{P}_{\boldsymbol{\Psi},p}(U_2; \boldsymbol{w}_0) = \mathcal{P}_{\boldsymbol{\Psi},p}(U_2; \mathcal{P}_{\boldsymbol{\Psi},p}(U_1; \boldsymbol{w}_0)) \tag{30}$$

Proof: Let $\boldsymbol{w}_1 = \mathcal{P}_{\boldsymbol{\Psi},p}(U_1; \boldsymbol{w}_0)$ and $\boldsymbol{w}_2 = \mathcal{P}_{\boldsymbol{\Psi},p}(U_2; \boldsymbol{w}_1)$. We have the following string of inequalities which hold for every $\boldsymbol{u} \in U_2$,

$$D(\boldsymbol{w}_1, \boldsymbol{w}_0) = D(\boldsymbol{u}, \boldsymbol{w}_0) - D(\boldsymbol{u}, \boldsymbol{w}_1), \tag{31}$$
$$D(\boldsymbol{w}_2, \boldsymbol{w}_1) = D(\boldsymbol{u}, \boldsymbol{w}_1) - D(\boldsymbol{u}, \boldsymbol{w}_2), \tag{32}$$
$$D(\boldsymbol{w}_1, \boldsymbol{w}_0) + D(\boldsymbol{w}_2, \boldsymbol{w}_1) = D(\boldsymbol{u}, \boldsymbol{w}_0) - D(\boldsymbol{u}, \boldsymbol{w}_2), \tag{33}$$
$$D(\boldsymbol{w}_2, \boldsymbol{w}_0) = D(\boldsymbol{u}, \boldsymbol{w}_0) - D(\boldsymbol{u}, \boldsymbol{w}_2), \tag{34}$$

where equations (31) and (32) follow from the pythagorean theorem (Lemma 13) equation (34) then follows from setting $\boldsymbol{u} = \boldsymbol{w}_2$ in (31) then substituting into (33). Equation (34) implies $\boldsymbol{w}_2$ is the projection of $\boldsymbol{w}_0$ onto U_2. ■

Since $\boldsymbol{\nabla}||\boldsymbol{u}||_{\boldsymbol{\Psi},p}^2|_{\boldsymbol{u}=\boldsymbol{0}} = \boldsymbol{0}$, we have the following corollary.

Corollary 17 *If $\boldsymbol{w}_0 := \boldsymbol{0}$ and we recursively define*

$$\boldsymbol{w}_{t+1} := \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \{D_{\boldsymbol{\Psi},p}(\boldsymbol{u}, \boldsymbol{w}_t) : u_{i_s} = y_s \ \forall s \leq t\}.$$

then

$$\boldsymbol{w}_\ell = \operatorname*{argmin}_{\boldsymbol{u} \in \mathbb{R}^n} \{\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p} : u_{i_s} = y_{i_s} \ \forall s \leq \ell\}$$

A.3 Proof of Theorem 1

In Corollary 17 we noted that the minimum $(\boldsymbol{\Psi}, p)$-seminorm interpolation algorithm is identical to a successive Bregman projection algorithm. We prove a bound for the latter. Let $\boldsymbol{u} \in \mathbb{R}^n$ be such that $u_{i_t} = y_t$ for all trials $t \leq \ell$. From (23) we have

$$\sum_{t=1}^{\ell} D_{\boldsymbol{\Psi},p}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) = D_{\boldsymbol{\Psi},p}(\boldsymbol{u}, \boldsymbol{w}_1) - D_{\boldsymbol{\Psi},p}(\boldsymbol{u}, \boldsymbol{w}_{\ell+1}) \tag{35}$$

Using Lemma 15 we lower bound $D_p(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t)$

$$(p-1)\|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_{\boldsymbol{\Psi},p}^2 \leq D_{\boldsymbol{\Psi},p}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t). \tag{36}$$

Note that there is a mistake, by convention, on the first trial since $\boldsymbol{w}_1 = \boldsymbol{0}$. Now, for each mistaken trial $t \in \mathcal{M}$ with $t \geq 2$, recalling Section 2 we define the linear functional $Z_t = E_{i_t} - E_{\eta_{i_t}}$, where

$$\eta_{i_t} = \operatorname*{argmin}_{i_s} \{\|E_{i_t} - E_{i_s}\|_{\boldsymbol{\Psi},p}^* : s \in \mathcal{M}, s < t\},$$

so that

$$\begin{aligned} 1 &\leq |Z_t(\boldsymbol{w}_{t+1}) - Z_t(\boldsymbol{w}_t)| && t \geq 2 \\ &= |Z_t(\boldsymbol{w}_{t+1} - \boldsymbol{w}_t)| && t \geq 2 \\ &\leq \|Z_t\|_{\boldsymbol{\Psi},p}^* \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_{\boldsymbol{\Psi},p} && t \geq 2 \\ &\leq \|Z_t\|_{\boldsymbol{\Psi},p}^{*^2} \|\boldsymbol{w}_{t+1} - \boldsymbol{w}_t\|_{\boldsymbol{\Psi},p}^2 && t \geq 2 \end{aligned} \tag{37}$$

thus on a mistaken trial $t \geq 2$ combining (36) and (37) gives

$$\frac{p-1}{\|Z_t\|_{\boldsymbol{\Psi},p}^{*^2}} \leq D_p(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \qquad t \geq 2\,. \tag{38}$$

We follow a technique introduced in [9]. Recalling Section 2, consider any cover $\mathcal{C} = \cup_k X_k$ which covers $X = \{i_1, i_2, \ldots, i_\ell\}$ with regard to the distance

$$d_{\boldsymbol{\Psi},p}(i,j) := \|E_i - E_j\|_{\boldsymbol{\Psi},p}^*\,,$$

with $\mathcal{N}(X, \rho, d_{\boldsymbol{\Psi},p})$ covering sets of diameter no greater than ρ. Let $\mathcal{F}$ be the set of trials in which a mistake first occurred on a cover set $\mathcal{F} = \cup_k \{\min\{t : i_t \in X_k\}\}$. Setting $\boldsymbol{w}_1 = \boldsymbol{0}$ we deduce from (35) and (38)

$$\begin{aligned} \sum_{t \in \mathcal{M}\setminus\mathcal{F}} \frac{1}{\|Z_t\|_{\boldsymbol{\Psi},p}^{*^2}} &\leq \sum_{t \in \mathcal{M}\setminus\{1\}} \frac{1}{\|Z_t\|_{\boldsymbol{\Psi},p}^{*^2}} \\ &\leq \frac{1}{p-1} \sum_{t \in \mathcal{M}\setminus\{1\}} D_{\boldsymbol{\Psi},p}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \\ &\leq \frac{1}{p-1} \sum_{t=1}^{\ell} D_{\boldsymbol{\Psi},p}(\boldsymbol{w}_{t+1}, \boldsymbol{w}_t) \\ &\leq \frac{\|\boldsymbol{u}\|_{\boldsymbol{\Psi},p}^2}{p-1} \end{aligned}$$

Recall that

$$\|Z_t\|_{\boldsymbol{\Psi},p}^* = d_{\boldsymbol{\Psi},p}(i_t, \eta_{i_t})\,.$$

Hence for any $t \in \mathcal{M}\setminus\mathcal{F}$ we have $\|Z_t\|_{\boldsymbol{\Psi},p}^* \leq \rho$. Hence as $|\mathcal{F}| \leq \mathcal{N}(X, \rho, d_{\boldsymbol{\Psi},p})$

$$\sum_{t \in \mathcal{M}-\mathcal{F}} 1 \leq \frac{\rho^2 \|\boldsymbol{u}\|_{\boldsymbol{\Psi},p}^2}{p-1}$$

$$|\mathcal{M}| \leq \mathcal{N}(X, \rho, d_{\boldsymbol{\Psi},p}) + \frac{\rho^2 \|\boldsymbol{u}\|_{\boldsymbol{\Psi},p}^2}{p-1}.$$

■