Relating Function Class Complexity and Cluster Structure in the Function Domain with Applications to Transduction

Guy Lever
University College London
g.lever@cs.ucl.ac.uk

Abstract

We relate function class complexity to structure in the function domain. This facilitates risk analysis relative to cluster structure in the input space which is particularly effective in semi-supervised learning. In particular we quantify the complexity of function classes defined over a graph in terms of the graph structure.

1 INTRODUCTION

We relate the learning process to cluster structure in the data which the learner is attempting to classify. It is well-known that data-dependent measures of function class complexity can lead to sharper risk bounds than those which do not capture the data distribution. We elaborate this principle by demonstrating a relationship between the richness of a function class and structural features in data drawn from the underlying input space $\mathcal{X}$ on which it acts. Specifically, a typical assumption in machine learning is that data are clustered and we refine a recent upper bound on Rademacher complexity of a function class, by relating it to cluster structure in the domain.

The intended application of these ideas is in the settings of transductive and semi-supervised learning. In (Chapelle and Zien, 2005) it is argued that virtually all successful semi-supervised learning techniques exploit the cluster assumption. In these frameworks we typically work with *empirically defined hypothesis classes* and it is natural to relate the learning process to the data which informs their construction. In such frameworks, an empirical metric on $\mathcal{X}$ which captures the intrinsic geometry of the data can be constructed giving an opportunity to relate learning to the *intrinsic* structure of data. A typical empirical metric, equivalent to electrical resistance distance, is particularly sensitive to clustering, thus relating function class complexity to the cluster structure of $\mathcal{X}$ is effective in this case.

A key object in these settings is a graph formed using the available data and, as pointed out in (Hanneke, 2006) it is important to reach an understanding of which properties of a graph are relevant to the performance of an algorithm which predicts the labeling of the graph, and we provide a further step in that direction: in the spirit of the work of (Herbster, 2008) in the online setting, we present risk bounds (and suggest a regularization algorithm) derived from the cluster structure of the graph in the resistance metric. In particular we bound the richness of a class of functions with bounded cut defined over the vertices of a graph. When a graph exhibits good k-means clustering, in the resistance metric, this cluster structure seems to serve as a sharp practical measure of the richness of classifiers over a graph when learning under the typical "smoothness" assumption of a small graph cut; this is intuitive and is established using a duality theory.

We finally give a semi-supervised risk bound in which the complexity terms are related to the cluster structure of the (labeled and unlabeled) data instances.

2 PRELIMINARIES

We denote by $\mathcal{H}$ a class of real-valued functions (*hypotheses*) mapping a domain $\mathcal{X}$ to a decision space $\mathcal{D}$. It is typical to assign a measure of complexity $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$ over functions in $\mathcal{H}$. This generally captures a prior belief that the hypothesis most likely to explain the relationship between data and their classification is simple, or that the true classifier respects the structure of the input space. Given $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$ we denote

$$\mathcal{H}_\alpha := \{h \in \mathcal{H} \ : \ F(h) \leq \alpha\}.$$

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9. Copyright 2010 by the authors.

We consider only function classes consisting of linear functions (in some, possibly kernelized, space) so that (soft) classification is $h(\boldsymbol{x}) = \langle h, \boldsymbol{x} \rangle$.

Given a distribution P_{XY} over the labeled input space $\mathcal{X} \times \mathcal{Y}$, and a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ we denote the true risk of $h \in \mathcal{H}$ by $\text{risk}^{\ell}(h) := \mathbb{E}_{(X,Y)\sim P_{XY}} \ell(h(X), Y)$, and the risk on a specific set $\mathcal{T}$ by $\text{risk}^{\ell}_{\mathcal{T}}(h) := \frac{1}{|\mathcal{T}|}\sum_{(x,y)\in\mathcal{T}} \ell(h(x), y)$ and, in particular, the empirical risk on a labeled training sample $\mathcal{S}$ by $\widehat{\text{risk}}^{\ell}_{\mathcal{S}}(h) := \frac{1}{|\mathcal{S}|}\sum_{(x,y)\in\mathcal{S}} \ell(h(x), y)$. When $\ell(\cdot,\cdot)$ is the $0-1$ loss of binary classification, $\ell_{0-1}(y, y') := \begin{cases} 0 \text{ if } y = y' \\ 1 \text{ if } y \neq y' \end{cases}$, then, for simplicity, we denote the corresponding binary classification risk and its empirical counterpart by $\text{risk}(\cdot)$ and $\widehat{\text{risk}}_{\mathcal{S}}(\cdot)$ respectively.

Definition The *empirical Rademacher complexity* of a function class $\mathcal{H}$, on a sample $\mathcal{S} = \{\boldsymbol{x}_1, ...\boldsymbol{x}_m\}$ is defined,

$$\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}) := \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^{m} h(\boldsymbol{x}_i)\sigma_i\right)\right]$$

where the σ_i are Rademacher random variables.

Definition Given a probability distribution over the draw of training samples from $\mathcal{X}$, the *Rademacher complexity* of a function class $\mathcal{H}$, w.r.t. samples of size m, is defined $\mathcal{R}_m(\mathcal{H}) := \mathbb{E}_{\mathcal{S}}(\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}))$.

Rademacher complexity provides generalization bounds which are typically sharper than VC bounds, since it captures the distribution of the data under consideration:

Theorem 2.1. *(Bartlett and Mendelson, 2002) Assume a loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is K-Lipschitz in its first argument and bounded by C, then for any $\delta > 0$, we have, with probability at least $1 - \delta$ over the draw of a training sample $\mathcal{S}$ of size m, that*

$$\sup_{h\in\mathcal{H}}\left(\text{risk}^{\ell}(h) - \widehat{\text{risk}}^{\ell}_{\mathcal{S}}(h)\right) \leq 2K\mathcal{R}_m(\mathcal{H}) + C\sqrt{\frac{\log\frac{1}{\delta}}{2m}}.$$

3 RELATING COMPLEXITY TO DOMAIN STRUCTURE

Definition Given a set $\mathcal{S}$ of points drawn from a vector space $\mathcal{X}$ a *clustering* of $\mathcal{S}$ is any partition $\mathcal{C} = \{\mathcal{C}_1, ...\mathcal{C}_N\}$ of $\mathcal{S}$. Given a metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$, for each k we define the *center* of $\mathcal{C}_k$ by $\boldsymbol{c}_k := \text{argmin}_{\boldsymbol{x}\in\mathcal{X}} \sum_{\boldsymbol{x}'\in\mathcal{C}_k} d^2(\boldsymbol{x}', \boldsymbol{x})$ and note that if $d(\cdot,\cdot)$ arises from an inner product, $d^2(\boldsymbol{x}, \boldsymbol{x}') = \langle \boldsymbol{x} - \boldsymbol{x}', \boldsymbol{x} - \boldsymbol{x}' \rangle$, then this is identical to the *centroid* $\boldsymbol{c}_k = \frac{1}{|\mathcal{C}_k|}\sum_{\boldsymbol{x}\in\mathcal{C}_k} \boldsymbol{x}$. For each $\boldsymbol{x} \in \mathcal{S}$ we denote its corresponding center by $c(\boldsymbol{x}) := \boldsymbol{c}_k$ where k is such that $\boldsymbol{x} \in \mathcal{C}_k$.

3.1 A "DUALITY" OF COMPLEXITY ON $\mathcal{H}$ AND DISTANCE ON $\mathcal{X}$

Given a class of linear functions $\mathcal{H} : \mathcal{X} \to \mathbb{R}$, any norm $||\cdot||$ on $\mathcal{H}$ (which would generally capture complexity in $\mathcal{H}$) gives rise to a specific metric $d(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geq 0}$ defined, via the dual norm $||\cdot||^*$, by

$$\begin{aligned} d(\boldsymbol{x}_i, \boldsymbol{x}_j) :&= ||\boldsymbol{x}_i - \boldsymbol{x}_j||^* \\ &= \sup_{h\in\mathcal{H}} \frac{|h(\boldsymbol{x}_i) - h(\boldsymbol{x}_j)|}{||h||}. \end{aligned}$$

Call such a metric the *implied metric*. Intuitively, if $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ can be classified differently by some simple hypothesis in h they are distant in $d(\cdot,\cdot)$, and conversely if they are distinctly classified only by complex hypotheses then they are close. Given a norm on $\mathcal{H}$, it is this implied metric which we use to quantify cluster structure in $\mathcal{X}$.

3.1.1 Examples

1. **Linear classification in an arbitrary RKHS.** Given any kernel K on a space $\mathcal{X}$, consider the reproducing kernel Hilbert space $\mathcal{H} = \overline{\text{span}\{K(\boldsymbol{x}, \cdot) : \boldsymbol{x} \in \mathcal{X}\}}$, consisting of all linear combinations of the *features* $\{K(\boldsymbol{x}, \cdot)\}_{\boldsymbol{x}\in\mathcal{X}}$, with inner product $\langle K(\boldsymbol{x}, \cdot), K(\boldsymbol{x}', \cdot)\rangle_{\mathcal{H}} := K(\boldsymbol{x}, \boldsymbol{x}')$, such that a given point $\boldsymbol{x} \in \mathcal{X}$ receives the (soft) classification $h(\boldsymbol{x}) = \langle h, K(\boldsymbol{x}, \cdot)\rangle_{\mathcal{H}}$. Kernel methods often amount to seeking a classifier by minimizing, or regularizing in $\mathcal{H}$ w.r.t., the norm $||h||_{\mathcal{H}} = \sqrt{\langle h, h\rangle_{\mathcal{H}}}$, whose dual, by the arguments above, defines an implied metric on the feature space (and by extension on $\mathcal{X}$),

$$\begin{aligned} d_K(\boldsymbol{x}, \boldsymbol{x}') :&= d(K(\boldsymbol{x}, \cdot), K(\boldsymbol{x}', \cdot)) \\ &= \sup_{||h||_{\mathcal{H}}\neq 0}\left\{\frac{|\langle h, K(\boldsymbol{x}, \cdot) - K(\boldsymbol{x}', \cdot)\rangle_{\mathcal{H}}|}{||h||_{\mathcal{H}}}\right\} \\ &= \sqrt{K(\boldsymbol{x}, \boldsymbol{x}) + K(\boldsymbol{x}', \boldsymbol{x}') - 2K(\boldsymbol{x}, \boldsymbol{x}')}. \end{aligned}$$

2. **Transductive classification on a graph.** Given an n-vertex connected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with (weighted) adjacency $\boldsymbol{A}$, we seek a classifier $\boldsymbol{h} \in \mathbb{R}^n$ which classifies the vertices $\mathcal{V} = \{v_1, ...v_n\}$ according to $\boldsymbol{h}(v_i) := \text{sgn}(\boldsymbol{h}^{\top}\boldsymbol{e}_i) = \text{sgn}(h_i)$, where we have identified each vertex v_i with the corresponding standard basis vector $\boldsymbol{e}_i$ in $\mathbb{R}^n$. A typical scheme is to minimize a *smooth-*

ness functional

$$F_{\boldsymbol{L}}(h) := \frac{1}{2}||\boldsymbol{h}||_{\boldsymbol{L}}^2 := \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{L}\boldsymbol{h} = \frac{1}{2}\sum_{(i,j)\in\mathcal{E}}(h_i - h_j)^2 A_{ij}$$

induced by the graph Laplacian $\boldsymbol{L}$, subject to label constraints (Zhu et al., 2003; Belkin et al., 2004). By following the above procedure, the dual of the semi-norm $||\boldsymbol{h}||_{\boldsymbol{L}}$, again implies a metric $d_{\boldsymbol{L}}(\cdot,\cdot)$ on $\mathcal{V}$ as follows,

$$\begin{aligned} d_{\boldsymbol{L}}(v_i, v_j) &:= ||\boldsymbol{e}_i - \boldsymbol{e}_j||_{\boldsymbol{L}}^* \\ &= \sup_{\boldsymbol{h}\in\mathbb{R}^n, ||\boldsymbol{h}||_{\boldsymbol{L}}\neq 0}\left\{\frac{|\boldsymbol{h}^\top(\boldsymbol{e}_i - \boldsymbol{e}_j)|}{||\boldsymbol{h}||_{\boldsymbol{L}}}\right\} \\ &= \sup_{\boldsymbol{h}\in\mathbb{R}^n, ||\boldsymbol{h}||_{\boldsymbol{L}}\neq 0}\left\{\frac{|(\boldsymbol{L}\boldsymbol{h})^\top\boldsymbol{L}^+(\boldsymbol{e}_i - \boldsymbol{e}_j)|}{\sqrt{(\boldsymbol{L}\boldsymbol{h})^\top\boldsymbol{L}^+(\boldsymbol{L}\boldsymbol{h})}}\right\} \\ &= \sup_{\boldsymbol{w}\in\mathrm{col}(\boldsymbol{L})}\left\{\frac{|\boldsymbol{w}^\top\boldsymbol{L}^+(\boldsymbol{e}_i - \boldsymbol{e}_j)|}{\sqrt{\boldsymbol{w}^\top\boldsymbol{L}^+\boldsymbol{w}}}\right\} \\ &= \sqrt{(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top\boldsymbol{L}^+(\boldsymbol{e}_i - \boldsymbol{e}_j)}, \end{aligned}$$

where $\boldsymbol{L}^+$ is the pseudoinverse of the graph Laplacian. This metric is equal to the square root of the electrical resistance between vertices on $\mathcal{G}$ (Klein and Randić, 1993), which arises by viewing the graph as an electrical network in which each edge corresponds to a resistor with conductance equal to the edge weight. This captures the geometry of a discrete input space by measuring the ease with which current flows through a body defined by the data which should be more appropriate than a generic distance in an ambient space.

3. **Semi-supervised classification.** The previous transductive example can be extended "out of sample" using arguments in (Sindhwani et al., 2005). We wish to build a classifier $h : \mathcal{X} \to \mathbb{R}$ and are given a sample of data points $\mathcal{S} = \{\boldsymbol{x}_1, ...\boldsymbol{x}_n\}$ from $\mathcal{X}$, but the true distribution of data from $\mathcal{X}$ is otherwise unknown. Given a kernel $K : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ which defines a RKHS of functions $\mathcal{H}$ over $\mathcal{X}$ with inner product $\langle\cdot,\cdot\rangle_{\mathcal{H}}$, we may consider the space $\widetilde{\mathcal{H}}$ of functions from $\mathcal{H}$ with modified inner product,

$$\langle h, g\rangle_{\widetilde{\mathcal{H}}} := \gamma_{\mathcal{H}}\langle h, g\rangle_{\mathcal{H}} + \gamma_{\mathcal{S}}\langle Sh, Sg\rangle_{\mathcal{S}},$$

where $S(\cdot)$ is the (linear) *point evaluation function* on $\mathcal{S}$, $Sh = (h(\boldsymbol{x}_1), ...h(\boldsymbol{x}_n))^\top$, and $\langle\cdot,\cdot\rangle_{\mathcal{S}}$ is an inner product over the space of functions over $\mathcal{S}$, and $\gamma_{\mathcal{H}}$, $\gamma_{\mathcal{S}}$ control the relative weight given to the inner product in $\mathcal{H}$ and the empirical inner product. If $\langle Sh, Sg\rangle_{\mathcal{S}} = (Sh)^\top\boldsymbol{M}(Sg)$, where $\boldsymbol{M}$ is a positive semi-definite matrix measuring smoothness on a graph $\mathcal{G}$ formed on $\mathcal{S}$, according to (Sindhwani et al., 2005) $\widetilde{\mathcal{H}}$ is a RKHS with kernel $\widetilde{K} : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$ given by

$$\widetilde{K}(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{\gamma_{\mathcal{H}}}K(\boldsymbol{x}, \boldsymbol{x}') - \frac{\gamma_{\mathcal{S}}}{\gamma_{\mathcal{H}}}\boldsymbol{k}_{\boldsymbol{x}}^\top(\gamma_{\mathcal{H}}\boldsymbol{I} + \gamma_{\mathcal{S}}\boldsymbol{M}\boldsymbol{K})^{-1}\boldsymbol{M}\boldsymbol{k}_{\boldsymbol{x}'}, \quad (1)$$

where $\boldsymbol{k}_{\boldsymbol{x}} = (K(\boldsymbol{x}_1, \boldsymbol{x}), ...K(\boldsymbol{x}_n, \boldsymbol{x}))^\top$, and $\boldsymbol{K}$ is the $n \times n$ Gram matrix $K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ for i, $j \leq n$.

The norm $||h||_{\widetilde{\mathcal{H}}} := \sqrt{\langle h, h\rangle_{\widetilde{\mathcal{H}}}}$ implies a metric on $\mathcal{X}$ given by

$$d_{\widetilde{K}}(\boldsymbol{x}, \boldsymbol{x}') = \sqrt{\widetilde{K}(\boldsymbol{x}, \boldsymbol{x}) + \widetilde{K}(\boldsymbol{x}', \boldsymbol{x}') - 2\widetilde{K}(\boldsymbol{x}, \boldsymbol{x}')},$$

thus, (an approximation to) the resistance distance (or another such empirical distance) can be extended to the whole of $\mathcal{X}$.

3.2 BOUNDING RADEMACHER COMPLEXITY

We require the notion of strong convexity (see e.g. Kakade et al., 2008).

Definition In an inner product space $\mathcal{U}$, a function $f : \mathcal{U} \to \mathbb{R}$ is *κ-strongly convex* w.r.t. a norm $||\cdot||$ on $\mathcal{U}$ if for all $\boldsymbol{u}, \boldsymbol{v} \in \mathcal{U}$ we have $f(\boldsymbol{u}) - f(\boldsymbol{v}) - \langle\boldsymbol{\nabla} f(\boldsymbol{v}), \boldsymbol{u} - \boldsymbol{v}\rangle \geq \frac{\kappa}{2}||\boldsymbol{u} - \boldsymbol{v}||^2$.

We require the following lemma, which is a generalization of (Kakade et al., 2008, Lemma 4).

Lemma 3.1. *Let S be a closed convex set and $F : S \to \mathbb{R}_{\geq 0}$ be κ-strongly convex w.r.t. a norm $||\cdot||$ over S. Let $\{Z_i\}_{i=1}^m$ be conditionally zero mean random variables such that $\mathbb{E}[(||Z_i||^*)^2] \leq r_i^2$. Then $\mathbb{E}[F^*(\sum_{i=1}^m Z_i)] \leq \frac{1}{2\kappa}\sum_{i=1}^m r_i^2$, where F^* denotes the Legendre-Fenchel conjugate of F.*

Proof Let $S_k := \sum_{i=1}^k Z_i$. F is κ-strongly convex w.r.t. $||\cdot||$ and so F^* is $\frac{1}{\kappa}$-strongly smooth w.r.t. $||\cdot||^*$ (Shalev-Shwartz, 2007), this means

$$F^*(S_{m-1} + Z_m) \leq F^*(S_{m-1}) + \langle\boldsymbol{\nabla} F^*(S_{m-1}), Z_m\rangle + \frac{1}{2\kappa}(||Z_m||^*)^2.$$

Denoting $\mathbb{E}_{k-1}(\cdot) := \mathbb{E}_{Z_k}(\cdot \mid Z_1, ...Z_{k-1})$ and taking conditional expectation gives,

$$\mathbb{E}_{m-1}[F^*(S_m)] \leq F^*(S_{m-1}) + \frac{1}{2\kappa}\mathbb{E}_{m-1}[(||Z_m||^*)^2],$$

and since $F^*(0) = \sup_{\boldsymbol{z}}(-F(\boldsymbol{z})) \leq 0$ the result follows by iterated use of the tower rule. □

We now refine a result of (Kakade et al., 2008, Theorem 3), which uses convex duality to bound Rademacher complexity, by accounting for cluster structure.

Theorem 3.2. *For a class $\mathcal{H}$ of bounded linear functions on a vector space $\mathcal{X}$, if $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$ is κ-strongly convex w.r.t. a norm $|| \cdot ||_F$ on $\mathcal{H}$, then for any sample $\mathcal{S} = \{\boldsymbol{x}_1, ...\boldsymbol{x}_m\}$ of points from $\mathcal{X}$ and all clusterings $\mathcal{C}$ of $\mathcal{S}$ we have, for all $\alpha > 0$,*

$$\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}_\alpha) \leq B\sqrt{\frac{|\mathcal{C}|}{m}} + \sqrt{\frac{2\alpha\rho_{\mathcal{S}}}{m\kappa}}, \tag{2}$$

where $\rho_{\mathcal{S}} := \frac{1}{m}\sum_{i=1}^m d_F^2(\boldsymbol{x}_i, c(\boldsymbol{x}_i))$, $d_F(\cdot,\cdot)$ is the implied metric on $\mathcal{X}$ and $B := \sup_{h\in\mathcal{H}_\alpha, \boldsymbol{x}\in\mathcal{X}} |h(\boldsymbol{x})|$. Further, for all clusterings $\mathcal{C}$ of $\mathcal{X}$ we have,

$$\mathcal{R}_m(\mathcal{H}_\alpha) \leq B\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}}\right] + \sqrt{\frac{2\alpha}{m\kappa}}\mathbb{E}_{\mathcal{S}}[\sqrt{\rho_{\mathcal{S}}}], \tag{3}$$

where expectation is over the draw of a random sample $\mathcal{S} = \{\boldsymbol{X}_1, ...\boldsymbol{X}_m\}$ from $\mathcal{X}$ and $\mathcal{C}_{\mathcal{S}} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{S} \cap \mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

Proof Let $\mathcal{C} = \{\mathcal{C}_1, ...\mathcal{C}_N\}$ be an arbitrary clustering of $\mathcal{S}$, and denote $m_j := |\mathcal{C}_j|$.

$$\begin{aligned}
\widehat{\mathcal{R}}_{\mathcal{S}}(\mathcal{H}_\alpha) &= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha} \langle h, \frac{1}{m}\sum_{i=1}^m \sigma_i \boldsymbol{x}_i\rangle\right] \\
&= \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha} \left(\langle h, \frac{1}{m}\sum_{i=1}^m \sigma_i c(\boldsymbol{x}_i)\rangle\right.\right. \\
&\quad \left.\left. + \langle h, \frac{1}{m}\sum_{i=1}^m \sigma_i(\boldsymbol{x}_i - c(\boldsymbol{x}_i))\rangle\right)\right] \\
&\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha} \langle h, \frac{1}{m}\sum_{j=1}^N \sum_{i:\boldsymbol{x}_i\in\mathcal{C}_j} \sigma_i \boldsymbol{c}_j\rangle\right] \\
&\quad + \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha} \langle h, \frac{1}{m}\sum_{i=1}^m \sigma_i(\boldsymbol{x}_i - c(\boldsymbol{x}_i))\rangle\right] \qquad (4)
\end{aligned}$$

We take these two terms in turn.

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha} \langle h, \frac{1}{m}\sum_{j=1}^N \sum_{i:\boldsymbol{x}_i\in\mathcal{C}_j} \sigma_i \boldsymbol{c}_j\rangle\right] \\
&\leq \frac{1}{m}\sum_{j=1}^N \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h_j\in\mathcal{H}_\alpha}\left(\left(\sum_{i:\boldsymbol{x}_i\in\mathcal{C}_j}\sigma_i\right)\langle h_j, \boldsymbol{c}_j\rangle\right)\right] \\
&\leq \frac{B}{m}\sum_{j=1}^N \mathbb{E}_{\boldsymbol{\sigma}}\left[\left|\sum_{i:\boldsymbol{x}_i\in\mathcal{C}_j}\sigma_i\right|\right] \\
&\leq \frac{B}{m}\sum_{j=1}^N \sqrt{m_j} \leq B\sqrt{\frac{N}{m}}. \qquad (5)
\end{aligned}$$

The final lines hold by the concavity of the square root and since $\sum_{j=1}^N m_j = m$. For the second term we follow the procedure in (Kakade et al., 2008): denote, $\boldsymbol{\theta} := \frac{1}{m}\sum_{i=1}^m \sigma_i(\boldsymbol{x}_i - c(\boldsymbol{x}_i))$. By Fenchel's inequality we have, for any $\lambda > 0$, $\langle h, \lambda\boldsymbol{\theta}\rangle \leq F(h) + F^*(\lambda\boldsymbol{\theta})$, so,

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha}\langle h, \boldsymbol{\theta}\rangle\right] &\leq \mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha}\left(\frac{F(h)}{\lambda}\right) + \frac{F^*(\lambda\boldsymbol{\theta})}{\lambda}\right] \\
&\leq \frac{\alpha}{\lambda} + \frac{1}{\lambda}\mathbb{E}_{\boldsymbol{\sigma}}\left[F^*(\lambda\boldsymbol{\theta})\right] \qquad (6)
\end{aligned}$$

We have that $||\frac{\lambda}{m}\sigma_i(\boldsymbol{x}_i - c(\boldsymbol{x}_i))||_F^* \leq \frac{\lambda d_F(\boldsymbol{x}_i, c(\boldsymbol{x}_i))}{m}$ and so by Lemma 3.1, $\mathbb{E}_{\boldsymbol{\sigma}}\left[F^*(\lambda\boldsymbol{\theta})\right] \leq \frac{\lambda^2}{2\kappa m^2}\sum_{i=1}^m (d_F(\boldsymbol{x}_i, c(\boldsymbol{x}_i)))^2 = \frac{\lambda^2\rho_{\mathcal{S}}}{2\kappa m}$. Therefore by picking $\lambda = \sqrt{\frac{2\alpha m\kappa}{\rho_{\mathcal{S}}}}$ in (6), we have,

$$\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{h\in\mathcal{H}_\alpha}\langle h, \boldsymbol{\theta}\rangle\right] \leq \sqrt{\frac{2\alpha\rho_{\mathcal{S}}}{m\kappa}}. \tag{7}$$

Combining (4), (5) and (7) gives the result. □

Note that these bounds are optimized by a good k-means clustering. In line with intuition, if the data distribution clusters and a good classifier respects this structure we can learn well with few examples and if the training sample reveals this structure we can be more confident in our risk analysis. In the apendix we suggest a risk analysis and algorithm derived from this result.

4 APPLICATION TO TRANSDUCTION

Statistical analyses of induction typically require that the hypothesis class is not informed by available data instances. Thus, being necessarily inherited from the geometry of the ambient representation space of the data, the metric in which structure is quantified in our theory is unlikely to ideally capture the intrinsic geometry of the data distribution. In the settings of transduction and semi-supervised learning the learner is more informed about the nature of the data distribution, reducing an element of uncertainty, and typically uses this information to choose a data-dependent hypothesis class implying a metric on the input space which captures the intrinsic geometry of the data. We will see that the empirically-defined metric implied on the input space by learning under typical "smoothness" assumptions is very sensitive to the clustering of data – much more so than any non-empirical metric.

Transduction refers to the learning framework in which the unlabeled instances from the test set are available at the start of the learning process, and it is assumed that they are drawn from the same underlying distribution, so that there is no bias in the labeling. For analytical purposes the setting is equivalently posed as follows: denote by $\mathcal{X}$ a finite input space and $\mathcal{Y}$ the corresponding label space so that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} = \{(\boldsymbol{x}_1, y_1), ...(\boldsymbol{x}_n, y_n)\}$ is the joint space of labeled inputs. From $\mathcal{Z}$ is drawn *uniformly without replacement* a *training sample* of labeled examples $\mathcal{S} = \{(\boldsymbol{X}_{s_1}, Y_{s_1}), ...(\boldsymbol{X}_{s_m}, Y_{s_m})\} \subseteq \mathcal{Z}$, leaving the remaining *test set* $\mathcal{T} = \{(\boldsymbol{X}_{t_1}, Y_{t_1}), ...(\boldsymbol{X}_{t_u}, Y_{t_u})\} = \mathcal{Z} \backslash \mathcal{S}$. The training sample together with all unlabeled instances from the test set $\{\boldsymbol{X}_{t_1}, ...\boldsymbol{X}_{t_u}\}$ is available to the learner and each unlabeled data point must be labeled. For a given loss function $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ a notion of risk suitable for a binary classifier $h : \mathcal{X} \to \mathcal{D}$ in this transductive setting is the average loss incurred on the test set, $\text{risk}^{\ell}_{\mathcal{T}}[h] := \frac{1}{|u|} \sum_{i=1}^{u} \ell(h(x_{t_i}), y_{t_i})$ which is sometimes called the *transductive risk.*

4.1 TRANSDUCTIVE RADEMACHER COMPLEXITY

Recalling Section 2, for clarity we henceforth denote the *transductive Rademacher complexity* by $\mathcal{R}^{\text{trs}}_m(\cdot)$ when the draw of a sample is uniform without replacement from a finite set and $\mathcal{R}^{\text{ind}}_m(\cdot)$ the standard inductive Rademacher complexity. We specialize the bound provided by (3) to the transductive setting by exploiting the concavity of $\sqrt{\cdot}$ and evaluating the expectation.

Corollary 4.1. *For a class $\mathcal{H}$ of bounded functions on a finite set $\mathcal{X}$, if $F : \mathcal{H} \to \mathbb{R}$ is κ-strongly convex w.r.t. a norm $||\cdot||_F$ on $\mathcal{H}$, then for all clusterings $\mathcal{C}$ of $\mathcal{X}$, for all $\alpha > 0$,*

$$\mathcal{R}^{\text{trs}}_m(\mathcal{H}_\alpha) \leq B \mathbb{E}_{\mathcal{S}} \left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}} \right] + \sqrt{\frac{2\alpha\rho}{m\kappa}}, \tag{8}$$

where $\rho := \frac{1}{n} \sum_{i=1}^{n} d^2_F(\boldsymbol{x}_i, c(\boldsymbol{x}_i))$, $d_F(\cdot, \cdot)$ denotes the implied metric on $\mathcal{X}$, $B := \sup_{\boldsymbol{h} \in \mathcal{H}_\alpha, \boldsymbol{x} \in \mathcal{X}} |h(\boldsymbol{x})|$, expectation is w.r.t. the (uniform without replacement) draw of a sample $\mathcal{S} = \{\boldsymbol{X}_{s_1}, ...\boldsymbol{X}_{s_m}\}$ from $\mathcal{X}$ and $\mathcal{C}_{\mathcal{S}} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{C}_k \cap \mathcal{S} \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

4.1.1 Binary Classifiers With Bounded Graph Cut

Transduction is typically posed as predicting the labeling of a partially labeled n-vertex graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. By representing each $v_i \in \mathcal{V}$ by the standard basis element $\boldsymbol{e}_i \in \mathbb{R}^n$ we seek a classifier $\boldsymbol{h} \in \mathcal{H}$, such that $\boldsymbol{h}(v_i) := h_i = \boldsymbol{h}^\top \boldsymbol{e}_i$ is the (soft) classification of vertex v_i. As discussed in Section 3.1.1 one principle involves minimizing the smoothness functional $F_{\boldsymbol{L}}(\boldsymbol{h}) := \frac{1}{2}\boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h}$, derived from the graph Laplacian. Note that for $\boldsymbol{h} \in \{-1, 1\}^n$, $\frac{1}{4}\boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h} = \text{cut}(\boldsymbol{h})$, the weighted sum of all edges connecting differently labeled vertices. This is 1-strongly convex w.r.t. $||\boldsymbol{h}||_{\boldsymbol{L}} := \sqrt{\boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h}}$ and the implied metric on $\mathcal{V}$ in this case is given by $d_{\boldsymbol{L}}(v_i, v_j) = \sqrt{(\boldsymbol{e}_i - \boldsymbol{e}_j)^\top \boldsymbol{L}^+ (\boldsymbol{e}_i - \boldsymbol{e}_j)}$, the square root of the electrical resistance on the graph. The above result therefore bounds the Rademacher complexity of the class

$$\mathcal{H}_\phi := \{\boldsymbol{h} \in \{-1, 1\}^n \ : \ \boldsymbol{h}^\top \boldsymbol{L} \boldsymbol{h} \leq \phi\}$$

of binary classifiers with bounded cut:

Corollary 4.2. *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, for any clustering $\mathcal{C}$ of $\mathcal{V}$, for all $\phi > 0$,*

$$\mathcal{R}^{\text{trs}}_m(\mathcal{H}_\phi) \leq \mathbb{E}_{\mathcal{S}} \left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}} \right] + \sqrt{\frac{\phi\rho}{m}}. \tag{9}$$

where $\rho := \frac{1}{n} \sum_{i=1}^{n} d^2_{\boldsymbol{L}}(v_i, c(v_i))$ and $\mathcal{C}_{\mathcal{S}} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{S} \cap \mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

Note that each centroid $c(v_i)$ is not a point on the graph but is represented in $\mathbb{R}^n$ by $\frac{1}{|\mathcal{C}_k|} \sum_{\{j : v_j \in \mathcal{C}_k\}} \boldsymbol{e}_j$ where k is such that $v_i \in \mathcal{C}_k$. Thus if $\mathcal{G}$ exhibits good k-means clustering in the (square root of the) resistance metric then the class of binary classifiers $\mathcal{H}_\phi$ is small. Because of the strong convexity framework we can also extend this analysis to the "p-resistances" of (Herbster and Lever, 2009), a generalization of p-norms to graphs.

Analysis For Prototypical Clusters

The prototypical example of a cluster is a clique, we consider the (unweighted) graph $\mathcal{K}$, a collection of N cliques $\mathcal{K}_1, ...\mathcal{K}_N$, such that $|\mathcal{K}_i| = k_i$, connected arbitrarily with edges (see Figure 1).

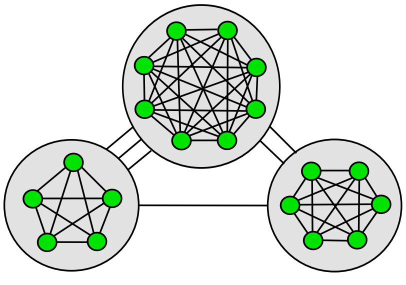

Figure 1: A Collection Of Cliques

By standard rules for resistors in series and parallel, the electrical resistance between any two distinct vertices in an k-clique is $\frac{2}{k}$, and, by Rayleigh's monotonicity principle, the inter-clique distances in a k-clique on $\mathcal{K}$ satisfy $d^2_{\boldsymbol{L}}(v_i, v_j) \leq \frac{2}{k}$. Now, for any set of n vertices

$\mathcal{V}'$ we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i:v_i\in\mathcal{V}'} d^2_{\boldsymbol{L}}(v_i, c(v_i)) \\
&= \frac{1}{2n^2}\sum_{i,j:v_i,v_j\in\mathcal{V}'} \left(\boldsymbol{e}_i-\boldsymbol{e}_j\right)^\top \boldsymbol{L}^+\left(\boldsymbol{e}_i-\boldsymbol{e}_j\right) \\
&\le \frac{1}{2}\frac{1}{n}\sum_{i:v_i\in\mathcal{V}'}\frac{1}{n-1}\sum_{j:v_j\in\mathcal{V}',j\neq i} d^2_{\boldsymbol{L}}(v_i,v_j),
\end{aligned}$$

so, on $\mathcal{K}$, the resistance distance from any vertex v_i to the centroid of its clique $\mathcal{K}_j$ satisfies $d^2_{\boldsymbol{L}}(v_i, c(v_i)) \le \frac{1}{k_j}$. Thus, for the graph $\mathcal{K}$, (9) implies that

$$\mathcal{R}^{\mathrm{trs}}_m(\mathcal{H}_\phi) \le \sqrt{\frac{N}{m}} + \sqrt{\frac{N\phi}{mn}}. \qquad (10)$$

Accounting for the cluster structure here offers significant improvement since the resistance distance between vertices in separate cliques is much larger (and on weighted graphs can be arbitrarily large).

Comparison To VC-Dimension Bounds

We now compare the result (9) to the bound of (Kleinberg et al., 2004) on the VC-dimension of $\mathcal{H}_\phi$ for unweighted graphs:

$$\mathrm{VCdim}(\mathcal{H}_\phi) = \mathcal{O}\left(\frac{\phi}{\phi^\star}\right), \qquad (11)$$

where $\phi^\star$ is the minimum number of edges that must be removed in order to disconnect the graph. Since $\mathcal{R}_m(\mathcal{H}) = \mathcal{O}\left(\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}\right)$, $\mathcal{R}_m(\mathcal{H})$ should be directly compared to $\sqrt{\frac{\mathrm{VCdim}(\mathcal{H})}{m}}$. For an unweighted collection of cliques $\mathcal{K}$ which is fairly easily disconnected[1], e.g. $\phi^\star < \frac{n}{N}$, the bound (10) can be preferred to $\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}_\phi)}{m}} = \mathcal{O}\left(\sqrt{\frac{\phi}{m\phi^\star}}\right)$ for ϕ reasonably large, e.g. $\phi > N\phi^\star$. We note that because of the appearance of the $\sqrt{\frac{1}{n}}$ term in the bound (10) there is a lot of slack to relax the connectivity of the graph while still maintaining a good bound. We note however that at the other end of the connectivity spectrum the bound (9) degrades: for example, for an unweighted path graph (9) becomes vacuous, at least for small m, and the VC bound is tight. This situation is improved by passing to p-resistances (Herbster and Lever, 2009): essentially the bound (8) holds simultaneously over a family of p-norms defined on the graph labellings and p-resistance, for $p \to 1$, is more suitable when the graph is sparse and partially solves the problem encountered here. Due to lack of space this will be presented in future work. We note however that (9) degrades here because $d^2_{\boldsymbol{L}}(v_i, c(v_i)) = \mathcal{O}(n)$ for a path graph, and this situation is probably far from typical in applications.

4.2 TRANSDUCTIVE RISK ANALYSIS

The following risk bound, due to (Pelckmans and Suykens, 2007) but generalized here, is valid in the transductive setting[2]. For completeness a proof is supplied in the appendix.

Theorem 4.3. *(Pelckmans and Suykens, 2007) For a given loss function $\ell : \mathcal{D}\times\mathcal{Y} \to \mathbb{R}_{\ge 0}$, K-Lipschitz in its first argument, bounded by C, for any $\delta > 0$, simultaneously for all $h \in \mathcal{H}$,*

$$\begin{aligned}
\mathbb{P}_{\mathcal{S}}\Bigg(&\mathrm{risk}^{\ell}_{\mathcal{T}}(h) \le \widehat{\mathrm{risk}}^{\ell}_{\mathcal{S}}(h) + 2K\frac{m+u}{\max(m,u)}\mathcal{R}^{\mathrm{trs}}_{\min(m,u)}(\mathcal{H}) \\
&+ C\left(\frac{1}{m}+\frac{1}{u}\right)\sqrt{\frac{\min(m,u)}{2}\log\frac{1}{\delta}}\Bigg) \ge 1-\delta,
\end{aligned}$$

where probability is w.r.t. the (uniform, without replacement) draw of the training sample $\mathcal{S} = \{(\boldsymbol{X}_{s_1}, Y_{s_1}), ...(\boldsymbol{X}_{s_m}, Y_{s_m})\}$ from $\mathcal{Z}$ and $\mathcal{T}\cup\mathcal{S} = \mathcal{Z}$.

We specialize this to the case of predicting the binary labeling of a graph $\mathcal{G}$ and apply the bound (9). Let $\mathcal{H} = \{-1, 1\}^n$ and $F_{\boldsymbol{L}}(\boldsymbol{h}) = \frac{1}{2}\boldsymbol{h}^\top\boldsymbol{L}\boldsymbol{h}$ where $\boldsymbol{L}$ is the Laplacian of $\mathcal{G}$. For simplicity we suppose $m < u$. We have $\mathcal{D} = \mathcal{Y} = \{-1, 1\}$ and by choosing the $0-1$ loss, which is $\frac{1}{2}$-Lipschitz for this function class, and bounded by 1, we have the following result bounding transductive binary classification risk of any algorithm which produces a binary labeling of a graph:

Theorem 4.4. *Given a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ for any clustering $\mathcal{C}$ of $\mathcal{V}$, for any $\delta > 0$, with probability at least $1-\delta$ over the draw of $\mathcal{S}$, simultaneously for all $\boldsymbol{h} \in \{-1, 1\}^n$,*

$$\begin{aligned}
&\mathrm{risk}_{\mathcal{T}}(\boldsymbol{h}) - \widehat{\mathrm{risk}}_{\mathcal{S}}(\boldsymbol{h}) \\
&\le \frac{n}{u}\left(\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}}\right] + 2\sqrt{\frac{F'_{\boldsymbol{L}}(\boldsymbol{h})\rho}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right), \qquad (12)
\end{aligned}$$

where $\rho := \frac{1}{n}\sum_{i=1}^n d^2_{\boldsymbol{L}}(v_i, c(v_i))$, $F'_{\boldsymbol{L}}(\boldsymbol{h}) := \min_{r\in\{1,2,...\}}\max\left(\phi_r, 2\frac{r+1}{r}F_{\boldsymbol{L}}(\boldsymbol{h})\right)$, $\phi_r := \frac{r\log 2}{2\rho}$ and $\mathcal{C}_{\mathcal{S}} := \{\mathcal{C}_k \in \mathcal{C} : \mathcal{S}\cap\mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the sample $\mathcal{S}$.

Proof Define the stratification[3]: $\mathcal{H}^{(0)} = \{\}$ and, for $t \in \{1, 2, ...\}$, $\mathcal{H}^{(t)} = \mathcal{H}_{\phi_t}$. Theorem 4.3 implies that

[1] Note that $\phi^\star$ doesn't reveal much about graph structure and could realistically be as small as 1 in practical applications

[2] As $u \to \infty$ we recover the inductive bound of Theorem 2.1.

[3] This technique is similar to that employed in (Balcan and Blum, Theorem 12).

with probability at least $1 - \frac{\delta}{2^t}$ simultaneously for all $\boldsymbol{h} \in \mathcal{H}^{(t)} \backslash \mathcal{H}^{(t-1)}$ we have,

$$\begin{aligned}
\operatorname{risk}_{\mathcal{T}}(\boldsymbol{h}) - \widehat{\operatorname{risk}}_{\mathcal{S}}(\boldsymbol{h}) &\leq \frac{n}{u}\left(\mathcal{R}_m^{\mathrm{trs}}(\mathcal{H}_{\phi_t}) + \sqrt{\frac{\log\frac{2^t}{\delta}}{2m}}\right) \\
&\leq \frac{n}{u}\left(\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}}\right] + \sqrt{\frac{\phi_t\rho}{m}} + \sqrt{\frac{t\log 2}{2m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right) \\
&\leq \frac{n}{u}\left(\mathbb{E}_{\mathcal{S}}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{S}}|}{m}}\right] + 2\sqrt{\frac{\phi_t\rho}{m}} + \sqrt{\frac{\log\frac{1}{\delta}}{2m}}\right). \qquad (13)
\end{aligned}$$

Now noting that for $r \in \{1, 2, ...\}$, $\phi_t > \phi_r$ implies that $t \geq r + 1$ and $\phi_t \leq \frac{t}{t-1}\phi_{t-1} \leq \frac{r+1}{r}\phi_{t-1}$, so

$$\phi_t \leq \min_{r\in\{1,2,...\}} \max\left(\phi_r, \frac{r+1}{r}\phi_{t-1}\right) \leq F'_{\boldsymbol{L}}(\boldsymbol{h}) \qquad (14)$$

The result follows by combining (14) with (13) and applying the union bound over all $t \in \{1, 2, ...\}$. □

This bound also suggests an algorithm obtained by minimizing the bound simultaneously over classifiers and clusterings; essentially a Laplacian regularization whose regularization parameters are determined by the cluster structure of the graph.

4.2.1 Comparison

We compare Theorem 4.4 to similar bounds in the literature. The following bound[4] is provided in (Hanneke, 2006).

Theorem 4.5. *(Hanneke, 2006, Corollary 2) With probability at least $1 - \delta$ simultaneously for all $\boldsymbol{h} \in \{-1, 1\}^n$,*

$$\operatorname{risk}_{\mathcal{T}}(\boldsymbol{h}) \leq \widehat{\operatorname{risk}}_{\mathcal{S}}(\boldsymbol{h}) + \sqrt{\frac{n(u+1)}{u^2}\frac{\frac{F_{\boldsymbol{L}}(\boldsymbol{h})}{\phi^\star}\ln n + \ln\frac{2(QW+1)}{\delta}}{2m}} \qquad (15)$$

where $\phi^\star$ is the minimum number of edges that must be removed to disconnect the graph, $W := \sum_{(i,j)\in\mathcal{E}} A_{ij}$, where $\boldsymbol{A}$ is the (weighted) adjacency of $\mathcal{G}$, and Q is the smallest positive rational number such that $QA_{ij} \in \mathbb{Z}$ for all $(i, j) \in \mathcal{E}$.

Since this is essentially equivalent to a bound derived from a the VC-dimension bound (11), we note that (ignoring multiplicative constants) (12) will be preferred to (15) whenever the Rademacher complexity bound (9) is preferred to the VC-dimension bound (11), and we refer the reader to the discussion of that subject in Section 4.1.1: for clustered graphs which are fairly easily disconnected, (12) seems preferable, nevertheless (15) remains tighter for sparser graphs, such as a path graph.

The following result relates the cardinality of $\mathcal{H}_\phi$ and transductive classification risk to the spectrum $\{\lambda_i\}_{i=1}^n$ of the graph Laplacian:

Theorem 4.6. *(Pelckmans et al., 2007, Theorem 1 and Theorem 2) With probability at least $1 - \delta$,*

$$\sup_{\boldsymbol{h}\in\mathcal{H}_\phi} |\operatorname{risk}_{\mathcal{T}}(\boldsymbol{h}) - \widehat{\operatorname{risk}}_{\mathcal{S}}(\boldsymbol{h})| \leq \sqrt{\frac{2(n-m+1)}{nm}\log\frac{|\mathcal{H}_\phi|}{\delta}}$$

with $|\mathcal{H}_\phi| \leq \left(\frac{en}{n_\phi}\right)^{n_\phi}$ where $n_\phi := |\{\lambda_i : \lambda_i \leq \phi\}|$.

We compare these results with that given by (12). For the simple toy example given in Figure 1, $n_\phi = |\mathcal{V}|$ for $\phi \geq 3$ and so the bound on $|\mathcal{H}_\phi|$ is vacuous. For a practical comparison we consider the MNIST data set of hand-written digits (Lecun and Cortes) and form a 4-NN graph from 500 instances each of the digits "0" and "1". The two approaches to bounding the richness of $\mathcal{H}_\phi$ on this data set and graph are summarized in Table 1 (results are averaged over 5 randomly chosen sets of data). The (average) true labeling $\boldsymbol{y}$ has a cut of 8, and so $\boldsymbol{y}^\top \boldsymbol{L}\boldsymbol{y} = 32$.

Table 1: Practical Evaluation Of Complexity Bounds

ϕ	n_ϕ	$\lvert\mathcal{H}_\phi\rvert$ (Thm. 4.6)	$\mathcal{R}_m^{\mathrm{trs}}(\mathcal{H}_\phi)$ (Eq. (9))
10	902	$\left(\frac{1000e}{902}\right)^{902}$	$\frac{1}{\sqrt{m}}\left(\sqrt{2} + 0.57\sqrt{10}\right)$
25	1000	e^{1000}	$\frac{1}{\sqrt{m}}\left(\sqrt{2} + 0.57\sqrt{25}\right)$
50	1000	e^{1000}	$\frac{1}{\sqrt{m}}\left(\sqrt{2} + 0.57\sqrt{50}\right)$

A comparison of the consequent bounds given by Theorem 4.6 and Theorem 4.3 apparently demonstrate that the bound (9) on the Rademacher complexity of $\mathcal{H}_\phi$ yields a sharper quantification of the richness of $\mathcal{H}_\phi$ on this data. Note that n_ϕ tends to be very large.

5 APPLICATION TO SEMI-SUPERVISED LEARNING

The above ideas provide a semi-supervised bound in which the complexity is entirely related to the cluster structure of the data sample. We are given a set $\mathcal{S} = \{(\boldsymbol{X}_{s_1}, Y_{s_1}), ...(\boldsymbol{X}_{s_m}, Y_{s_m})\}$ of m labeled instances drawn i.i.d. from P_{XY} over $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ and a set $\mathcal{X}_{\mathcal{T}} = \{\boldsymbol{X}_{t_1}, ...\boldsymbol{X}_{t_u}\}$ of u unlabeled instances drawn

[4] We note that (Hanneke, 2006) provides a sharper implicit bound. Since we are interested in the essential dependence of these bounds on structural quantities of the graph we compare, for simplicity, to the explicit bound only.

i.i.d. from the marginal P_X. Let $\mathcal{X}_{\mathcal{S}} := \{\boldsymbol{X}_{s_1}, \ldots \boldsymbol{X}_{s_m}\}$ and $\mathcal{I} := \mathcal{X}_{\mathcal{T}} \cup \mathcal{X}_{\mathcal{S}}$ denote the set of all $n = m + u$ instances. Consider a space $\mathcal{H}$ of bounded hypotheses mapping $\mathcal{X}$ to $\mathcal{D}$ and a complexity measure $F : \mathcal{H} \to \mathbb{R}_{\geq 0}$, κ-strongly convex w.r.t. a norm $||\cdot||_F$ on $\mathcal{H}$ and which is not informed by the sample of data instances, and let $\mathcal{H}_\alpha := \{h \in \mathcal{H} \ : \ F(h) \leq \alpha\}$. We then consider the space $\widetilde{\mathcal{H}}$ of functions from $\mathcal{H}$ with "modified" complexity $\widetilde{F} : \mathcal{H} \to \mathbb{R}_{\geq 0}$, $\widetilde{\kappa}$-strongly convex w.r.t. a norm $||\cdot||_{\widetilde{F}}$ on $\mathcal{H}$, which can take into account an empirical complexity measure derived from the entire sample of instances $\mathcal{I}$. The following semi-supervised bound on hypotheses from the empirically defined $\widetilde{\mathcal{H}}_\beta := \{h \in \mathcal{H}_\alpha \ : \ \widetilde{F}(h) \leq \beta\}$, like the sample complexity result of (Balcan and Blum, 2005, Theorem 5) relies on the (non-empirical) $\mathcal{H}_\alpha$ to prove convergence of transductive to inductive risk. It is proved in the appendix. The idea here is that the terms relating to the (non-empirical) hypothesis space $\mathcal{H}_\alpha$ decay as $\mathcal{O}(\frac{1}{\sqrt{n}})$.

Theorem 5.1. *Let $\ell : \mathcal{D} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ be a loss function, K-Lipschitz in its first argument and bounded by C. Then simultaneously for all $h \in \widetilde{\mathcal{H}}_\beta$ we have,*

$$\mathbb{P}\Bigg(\mathrm{risk}^\ell(h) \leq \widehat{\mathrm{risk}}^\ell_{\mathcal{S}}(h) + 2K\mathcal{R}^{\mathrm{trs}}_m(\widetilde{\mathcal{H}}_\beta) + 2K\widehat{\mathcal{R}}^{\mathrm{ind}}_{\mathcal{I}}(\mathcal{H}_\alpha) + C\left(\sqrt{\frac{1}{2m}\log\frac{2}{\delta}} + 3\sqrt{\frac{1}{2n}\log\frac{4}{\delta}}\right)\Bigg) \geq 1 - \delta, \quad (16)$$

where probability is w.r.t. the draw of the labeled and unlabeled data from P_{XY}. Further, for all clusterings $\mathcal{C}, \mathcal{C}'$ of $\mathcal{I}$,

$$\mathcal{R}^{\mathrm{trs}}_m(\widetilde{\mathcal{H}}_\beta) \leq B\mathbb{E}\left[\sqrt{\frac{|\mathcal{C}_{\mathcal{X}_{\mathcal{S}}}|}{m}}\right] + \sqrt{\frac{2\beta}{mn\widetilde{\kappa}}\sum_{\boldsymbol{X}\in\mathcal{I}} d^2_{\widetilde{F}}(\boldsymbol{X}, c(\boldsymbol{X}))},$$

and,

$$\widehat{\mathcal{R}}^{\mathrm{ind}}_{\mathcal{I}}(\mathcal{H}_\alpha) \leq B\sqrt{\frac{|\mathcal{C}'|}{n}} + \frac{1}{n}\sqrt{\frac{2\alpha}{\kappa}\sum_{\boldsymbol{X}\in\mathcal{I}} d^2_F(\boldsymbol{X}, c'(\boldsymbol{X}))},$$

where $\mathcal{C}_{\mathcal{X}_{\mathcal{S}}} := \{\mathcal{C}_k \in \mathcal{C} \ : \ \mathcal{X}_{\mathcal{S}} \cap \mathcal{C}_k \neq \Phi\}$ is the clustering restricted to the labeled instances, expectation is with respect to the (uniform without replacement) draw of $\mathcal{X}_{\mathcal{S}}$ from $\mathcal{I}$, $d_F(\cdot,\cdot)$ and $d_{\widetilde{F}}(\cdot,\cdot)$ are the metrics on $\mathcal{X}$ implied by $||\cdot||_F$ and $||\cdot||_{\widetilde{F}}$, and $B := \sup_{h\in\mathcal{H}_\alpha, \boldsymbol{x}\in\mathcal{X}} |\langle h, \boldsymbol{x}\rangle|$.

Aknowledgements

I would like to thank Mark Herbster, John Shawe-Taylor and the anonymous reviewers for their insights.

References

M. Balcan and A. Blum. A discriminative model for semi-supervised learning. *JACM to appear.*

M.-F. Balcan and A. Blum. A pac-style model for learning from labeled and unlabeled data. In *COLT*, volume 3559. Springer, 2005.

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

S. Boucheron, O. Bousquet, and G. Lugosi. Theory of classification: A survey of some recent advances. volume 9, pages 323 – 375, 2005.

O. Chapelle and A. Zien. Semi-supervised classification by low density separation, 2005.

R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. In *COLT*, volume 4539, pages 157–171. Springer, 2007.

S. Hanneke. An analysis of graph cut size for transductive learning. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, 2006.

M. Herbster. Exploiting cluster-structure to predict the labeling of a graph. In *ALT*, volume 5254 of *Lecture Notes in Computer Science*, pages 54–69. Springer, 2008.

M. Herbster and G. Lever. Predicting the labelling of a graph via p-norm interpolation. In *COLT '09: Proceedings*, 2009.

S. M. Kakade, K. Sridharan, and A. Tewari. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *NIPS*. MIT Press, 2008.

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

J. M. Kleinberg, M. Sandler, and A. Slivkins. Network failure detection and graph connectivity. In *SODA*, pages 76–85. SIAM, 2004.

Y. Lecun and C. Cortes. The mnist database of handwritten digits. URL `http://yann.lecun.com/exdb/mnist/`.

R. Meir and T. Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

K. Pelckmans and J. A. K. Suykens. Transductive rademacher complexities for learning over a graph. In *MLG*, 2007.

K. Pelckmans, J. Shawe-Taylor, J. A. K. Suykens, and B. D. Moor. Margin based transductive graph cuts using linear programming. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, 2007.

S. Shalev-Shwartz. Online learning: Theory, algorithms and applications. In *PhD Thesis*. The Hebrew University, 2007.

V. Sindhwani, P. Niyogi, and M. Belkin. Beyond the point cloud: from transductive to semi-supervised learning. In *ICML '05: Proc. 22nd international conference on Machine learning*, 2005.

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.