

Transductive Rademacher Complexity and its Applications

Ran El-Yaniv and Dmitry Pechyony

Computer Science Department
Technion - Israel Institute of Technology
{rani,pechyony}@cs.technion.ac.il

Abstract. We present data-dependent error bounds for transductive learning based on transductive Rademacher complexity. For specific algorithms we provide bounds on their Rademacher complexity based on their “unlabeled-labeled” decomposition. This decomposition technique applies to many current and practical graph-based algorithms. Finally, we present a new PAC-Bayesian bound for mixtures of transductive algorithms based on our Rademacher bounds.

1 Introduction

Transductive learning was already proposed and briefly studied more than thirty years ago [19], but only lately has it been empirically recognized that transduction can often facilitate more efficient or accurate learning than the traditional supervised learning approach (see, e.g., [8]). This recognition has motivated a flurry of recent activity focusing on transductive learning, with many new algorithms and heuristics being proposed. Nevertheless, issues such as the identification of “universally” effective learning principles for transduction remain unresolved. Statistical learning theory provides a principled approach to attack such questions through the study of error bounds. For example, in inductive learning such bounds have proven instrumental in characterizing learning principles and deriving practical algorithms.

So far, several general error bounds for transductive inference have been developed [20, 6, 9, 12]. In this paper we continue this fruitful line of research and develop tight, high probability data-dependent error bounds for transduction based on the Rademacher complexity. Inspired by [16] (Theorem 24), our main result in this regard is Theorem 2, offering a sufficient condition for transductive learning. While this result is syntactically similar to known inductive Rademacher bounds (see, e.g., [3]), it is fundamentally different in the sense that the transductive Rademacher averages are taken with respect to hypothesis spaces that can depend on the unlabeled training and test examples. This opportunity is unavailable in the inductive setting where the hypothesis space must be fixed before any example is observed.

Our second contribution is a technique for establishing Rademacher bounds for specific algorithms based on their *unlabeled-labeled decomposition (ULD)*. In

this decomposition we present the algorithm as $\text{sgn}(K\alpha)$, where K is a matrix that depends on the unlabeled data and α is a vector that may depend on all given information including the labeled training set. We show that many state-of-the-art algorithms have non-trivial ULD leading to tight error bounds. In particular, we provide such bounds for the Gaussian random field transductive algorithm of [23], the “consistency method” of [22], the spectral graph transducer (SGT) algorithm of [15], the eigenmap algorithm of [5] and the Tikhonov regularization algorithm of [4].

We also show a simple Monte-Carlo scheme for bounding the Rademacher complexity of any transductive algorithm using its ULD. We demonstrate the efficacy of this scheme for the “consistency method” of [22]. Experimental evidence from [8] (Chapter 21) indicates that the SGT algorithm of [15] is amongst the better transductive algorithms currently known. Motivated by this fact we derived a specific error bound for this algorithm. Our final contribution is a PAC-Bayesian bound for transductive mixture algorithms. This result, which is stated in Theorem 3, is obtained as a consequence of Theorem 2 using the techniques of [17]. This result motivates the use of ensemble methods in transduction that are yet to be explored in this setting.

Related Work. Vapnik [20] presented the first general 0/1 loss bounds for transduction. His bounds are implicit in the sense that tail probabilities are specified in the bound as the outcome of a computational routine. Vapnik’s bounds can be refined to include prior “beliefs” as noted in [9]. Similar implicit but somewhat tighter bounds were developed in [6] for the 0/1 loss case. Explicit PAC-Bayesian transductive bounds for any bounded loss function were presented in [9]. The bounds of [1] for semi-supervised learning also hold in the transductive setting, making them conceptually similar to some transductive PAC-Bayesian bounds. General error bounds based on stability were developed in [12].

Effective applications of the general bounds mentioned above to particular algorithms or “learning principles” is not automatic. In the case of the PAC-Bayesian bounds several such successful applications are presented in terms of appropriate “priors” that promote various structural properties of the data [9, 11, 13]. Ad-hoc bounds for particular algorithms were developed in [4, 21].

Error bounds based on the Rademacher complexity are a well-established topic in induction (see [3] and references therein). The first Rademacher transductive risk bound was presented in [16]. This bound, which is a straightforward extension of the inductive Rademacher techniques of [3], is limited to the special case when training and test sets are of equal size. The bound presented here overcomes this limitation.

2 Transductive Rademacher complexity

We begin with some definitions. Consider a fixed set $S_{m+u} = (\langle x_i, y_i \rangle)_{i=1}^{m+u}$ of $m + u$ points x_i in some space together with their labels y_i . The learner is provided with the (unlabeled) *full-sample* $X_{m+u} = \{x_i\}_{i=1}^{m+u}$. A set consisting

of m points is selected from X_{m+u} uniformly at random among all subsets of size m . These m points together with their labels are given to the learner as a *training set*. Re-numbering the points we denote the training set points by $X_m = \{x_1, \dots, x_m\}$ and the labeled training set by $S_m = (\langle x_i, y_i \rangle)_{i=1}^m$. The set $X_u \triangleq \{x_{m+1}, \dots, x_{m+u}\} = X_{m+u} \setminus X_m$ is called the *test set*. The learner's goal is to predict the labels of the test points in X_u based on $S_m \cup X_u$.

This paper focuses on binary learning problems where labels $y \in \{\pm 1\}$. The learning algorithms we consider generate “soft classification” vectors $\mathbf{h} = (h(1), \dots, h(m+u)) \in \mathbb{R}^{m+u}$, where $h(i)$ (or $h(x_i)$) is the soft, or confidence-rated, label of example x_i given by the “hypothesis” \mathbf{h} . For actual (binary) classification of x_i the algorithm outputs $\text{sgn}(h(i))$.

Based on the full-sample X_{m+u} the algorithm selects an hypothesis space \mathcal{H} of such soft classification hypotheses. Then, given the labels of training points the algorithm selects one hypothesis from \mathcal{H} for classification. The goal is to minimize its *test error* $\mathcal{L}_u(\mathbf{h}) \triangleq \frac{1}{u} \sum_{i=m+1}^{m+u} \ell(h(x_i), y_i)$ w.r.t. the 0/1 loss function ℓ . In this work we use also the margin loss function ℓ_γ . For a positive real γ , $\ell_\gamma(y_1, y_2) = 0$ if $y_1 y_2 \geq \gamma$ and $\ell_\gamma(y_1, y_2) = \min\{1, 1 - y_1 y_2 / \gamma\}$ otherwise. The *empirical (margin) error* of \mathbf{h} is $\widehat{\mathcal{L}}_m^\gamma(\mathbf{h}) \triangleq \frac{1}{m} \sum_{i=1}^m \ell_\gamma(h(x_i), y_i)$. We denote by $\mathcal{L}_u^\gamma(\mathbf{h})$ the *test margin error*.

We adapt the inductive Rademacher complexity to our transductive setting but generalize it a bit to include “neutral” Rademacher values also.

Definition 1 (Transductive Rademacher Complexity). *Let $\mathcal{V} \subseteq \mathbb{R}^{m+u}$ and $p \in [0, 1/2]$. Let $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_{m+u})$ be a vector of i.i.d. random variables such that*

$$\sigma_i \triangleq \begin{cases} 1 & \text{w.p. } p; \\ -1 & \text{w.p. } p; \\ 0 & \text{w.p. } 1 - 2p. \end{cases} \quad (1)$$

The Transductive Rademacher Complexity with parameter p is $R_{m+u}(\mathcal{V}, p) \triangleq (\frac{1}{m} + \frac{1}{u}) \cdot \mathbf{E}_{\boldsymbol{\sigma}} \{\sup_{\mathbf{v} \in \mathcal{V}} \boldsymbol{\sigma} \cdot \mathbf{v}\}$.

For the case $p = 1/2$ and $m = u$ the resulting transductive complexity coincides with the standard inductive definition (see, e.g., [3]) up to the normalization factor $(\frac{1}{m} + \frac{1}{u})$. Whenever $p < 1/2$, some Rademacher variables will obtain (neutral) zero values and reduce the complexity (see Lemma 1). We use this parameterized version of the complexity to tighten our bounds. Notice that the transductive complexity is an empirical quantity that does not depend on any underlying distribution. Also, the transductive complexity depends on the test points whereas the inductive complexity only depends on the (unlabeled) training points.

The following lemma states that $R_{m+u}(\mathcal{V}, p)$ is monotone increasing with p . The proof of the lemma is omitted and will appear in the full version. The proof of Lemma 1 is based on the technique used in the proof of Lemma 5 in [17].

Lemma 1. *For any $\mathcal{V} \subseteq \mathbb{R}^{m+u}$ and $0 \leq p_1 < p_2 \leq 1/2$, $R_{m+u}(\mathcal{V}, p_1) < R_{m+u}(\mathcal{V}, p_2)$.*

The statements that follow utilize the Rademacher complexity with $p_0 \triangleq \frac{mu}{(m+u)^2}$. We abbreviate $R_{m+u}(\mathcal{V}) \triangleq R_{m+u}(\mathcal{V}, p_0)$. By Lemma 1, all our bounds apply also to $R_{m+u}(\mathcal{V}, p)$ for all $p > p_0$.

3 Uniform concentration inequality for a set of vectors

Denote by I_r^s for the set of natural numbers $\{r, \dots, s\}$ ($r < s$). Let $\mathbf{Z} \triangleq \mathbf{Z}_1^{m+u} \triangleq (Z_1, \dots, Z_{m+u})$ be a *random permutation vector* where the variable Z_k , $k \in I_1^{m+u}$, is the k th component of a permutation of I_1^{m+u} that is chosen uniformly at random. Let \mathbf{Z}^{ij} be a perturbed permutation vector obtained by exchanging Z_i and Z_j in \mathbf{Z} . Any function f on permutations of I_1^{m+u} is called *(m, u)-permutation symmetric* if $f(\mathbf{Z}) \triangleq f(Z_1, \dots, Z_{m+u})$ is symmetric on Z_1, \dots, Z_m as well as on Z_{m+1}, \dots, Z_{m+u} .

The following lemma (that will be utilized in the proof of Theorem 1) presents a concentration inequality that is a slight extension of Lemma 2 from [12]. The argument relies on the Hoeffding-Azuma inequality for martingales (the proof will appear in the full version). Note that a similar but weaker statement can be extracted using the technique of [16] (Claim 2 of the proof of Theorem 24).¹

Lemma 2 ([12]). *Let \mathbf{Z} be a random permutation vector over I_1^{m+u} . Let $f(\mathbf{Z})$ be an (m, u)-permutation symmetric function satisfying $|f(\mathbf{Z}) - f(\mathbf{Z}^{ij})| \leq \beta$ for all $i \in I_1^m$, $j \in I_{m+1}^{m+u}$. Then*

$$\mathbf{P}_{\mathbf{Z}} \{f(\mathbf{Z}) - \mathbf{E}_{\mathbf{Z}} \{f(\mathbf{Z})\} \geq \epsilon\} \leq \exp\left(-\frac{\epsilon^2(m+u)}{2\beta^2 mu}\right). \quad (2)$$

Let \mathcal{V} be a set of vectors in $[B_1, B_2]^{m+u}$, $B_1 \leq 0$, $B_2 \geq 0$ and set $B \triangleq B_2 - B_1$, $B_{\max} = \max(|B_1|, |B_2|)$. Consider two independent permutations of I_1^{m+u} , \mathbf{Z} and \mathbf{Z}' . For any $\mathbf{v} \in \mathcal{V}$ denote by $\mathbf{v}(\mathbf{Z}) \triangleq (v(Z_1), v(Z_2), \dots, v(Z_{m+u}))$ the vector \mathbf{v} permuted according to \mathbf{Z} . We use the following abbreviations for averages of \mathbf{v} over subsets of its components: $\mathbf{H}_k\{\mathbf{v}(\mathbf{Z})\} \triangleq \frac{1}{m} \sum_{i=1}^k v(Z_i)$, $\mathbf{T}_k\{\mathbf{v}(\mathbf{Z})\} \triangleq \frac{1}{u} \sum_{i=k+1}^{m+u} v(Z_i)$ (note that \mathbf{H} stands for ‘head’ and \mathbf{T} , for ‘tail’). In the special case where $k = m$ we set $\mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \triangleq \mathbf{H}_m\{\mathbf{v}(\mathbf{Z})\}$, and $\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} \triangleq \mathbf{T}_m\{\mathbf{v}(\mathbf{Z})\}$. Finally, the average component of \mathbf{v} is denoted $\bar{\mathbf{v}} \triangleq \frac{1}{m+u} \sum_{i=1}^{m+u} v(i)$.

¹ The idea in [16] is to represent a function of the permutation of $m+u$ indices as a function of independent random variables and use McDiarmid’s bounded difference inequality for *independent* random variables. It is not hard to extend the result of [16] for $m = u$ to the general case of $m \neq u$, but the resulting concentration inequality would have a $1/(m+u)$ term instead of the $(m+u)/(mu)$ term as in our Lemma 2. We achieve this advantage by exploiting the (m, u) -symmetry. The resulting sharper bound is critical for obtaining converging error bounds using our techniques.

For any $\mathbf{v} \in \mathcal{V}$ and any permutation \mathbf{Z} of I_1^{m+u} we have

$$\begin{aligned}
\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} &= \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \mathbf{T}\{\mathbf{v}(\mathbf{Z})\} - \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \\
&\leq \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \sup_{\mathbf{v} \in \mathcal{V}} \left[\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} - \bar{\mathbf{v}} + \bar{\mathbf{v}} - \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \right] \\
&= \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \sup_{\mathbf{v} \in \mathcal{V}} \left[\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} - \mathbf{E}_{\mathbf{Z}'} \mathbf{T}\{\mathbf{v}(\mathbf{Z}')\} + \mathbf{E}_{\mathbf{Z}'} \mathbf{H}\{\mathbf{v}(\mathbf{Z}')\} - \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \right] \\
&\leq \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \underbrace{\mathbf{E}_{\mathbf{Z}'} \sup_{\mathbf{v} \in \mathcal{V}} \left[\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} - \mathbf{T}\{\mathbf{v}(\mathbf{Z}')\} + \mathbf{H}\{\mathbf{v}(\mathbf{Z}')\} - \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \right]}_{\triangleq g(\mathbf{Z})}.
\end{aligned}$$

The function $g(\mathbf{Z})$ is (m, u) -permutation symmetric in \mathbf{Z} . It can be verified that $|g(\mathbf{Z}) - g(\mathbf{Z}^{ij})| \leq B \left(\frac{1}{m} + \frac{1}{u} \right)$. Therefore, we can apply Lemma 2 with $\beta \triangleq B \left(\frac{1}{m} + \frac{1}{u} \right)$ to $g(\mathbf{Z})$. Since $\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} - \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} \leq g(\mathbf{Z})$, we obtain, with probability of at least $1 - \delta$ over random permutation \mathbf{Z} of I_1^{m+u} , for all $\mathbf{v} \in \mathcal{V}$:

$$\begin{aligned}
\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} &\leq \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} + B \left(\frac{1}{m} + \frac{1}{u} \right) \sqrt{\frac{2mu}{m+u} \ln \frac{1}{\delta}} \\
&= \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + \mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} + B \sqrt{2 \left(\frac{1}{m} + \frac{1}{u} \right) \ln \frac{1}{\delta}}. \tag{3}
\end{aligned}$$

Our goal is to bound the expectation $\mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\}$. For technical convenience we use the following definition of the Rademacher complexity with pairwise Rademacher variables. This definition is equivalent to Def. 1 with $p = \frac{mu}{(m+u)^2}$.

Definition 2. Let $\mathbf{v} = (v(1), \dots, v(m+u)) \in \mathbb{R}^{m+u}$. Let \mathcal{V} be a set of vectors from \mathbb{R}^{m+u} . Let $\tilde{\sigma} = \{\tilde{\sigma}_i\}_{i=1}^{m+u}$ be a vector of i.i.d. random variables defined as:

$$\tilde{\sigma}_i = (\tilde{\sigma}_{i,1}, \tilde{\sigma}_{i,2}) = \begin{cases} \left(-\frac{1}{m}, -\frac{1}{u} \right) & \text{with prob. } \frac{mu}{(m+u)^2}, \\ \left(-\frac{1}{m}, \frac{1}{m} \right) & \text{with prob. } \frac{m^2}{(m+u)^2}, \\ \left(\frac{1}{u}, \frac{1}{m} \right) & \text{with prob. } \frac{mu}{(m+u)^2}, \\ \left(\frac{1}{u}, -\frac{1}{u} \right) & \text{with prob. } \frac{u^2}{(m+u)^2}. \end{cases} \tag{4}$$

The ‘‘pairwise’’ transductive Rademacher complexity is defined to be

$$\tilde{R}_{m+u}(\mathcal{V}) \triangleq \mathbf{E}_{\tilde{\sigma}} \left\{ \sup_{\mathbf{v} \in \mathcal{V}} \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) v(i) \right\}. \tag{5}$$

It is not hard to see from the definition of σ and $\tilde{\sigma}$ that $R_{m+u}(\mathcal{V}) = \tilde{R}_{m+u}(\mathcal{V})$.

Lemma 3. Let \mathbf{Z} be a random permutation of I_1^{m+u} . Let $c_0 = \sqrt{\frac{32 \ln(4e)}{3}} < 5.05$. Then

$$\mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} \leq \tilde{R}_{m+u}(\mathcal{V}) + c_0 B \left(\frac{1}{u} + \frac{1}{m} \right) \sqrt{\min(m, u)}. \tag{6}$$

Proof: The proof of Lemma 3 is based on ideas from the proof of Lemma 3 in [3]. Let n_1, n_2 and n_3 be the number of random variables $\tilde{\sigma}_i$ realizing the value $(-\frac{1}{m}, -\frac{1}{u}), (-\frac{1}{m}, \frac{1}{m}), (\frac{1}{u}, \frac{1}{m})$, respectively. Set $N_1 \triangleq n_1 + n_2$ and $N_2 \triangleq n_2 + n_3$. Note that the n_i 's and N_i 's are random variables. Denote by \mathbf{R} the distribution of $\tilde{\sigma}$ defined by (4) and by $\mathbf{R}(N_1, N_2)$, the distribution \mathbf{R} conditioned on the events $n_1 + n_2 = N_1$ and $n_2 + n_3 = N_2$. We define

$$s(N_1, N_2) \triangleq \mathbf{E}_{\tilde{\sigma} \sim \mathbf{R}(N_1, N_2)} \left\{ \sup_{\mathbf{v} \in \mathcal{V}} \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) v(i) \right\}. \quad (7)$$

The rest of the proof is based on the following three claims:

Claim 1. $\tilde{R}_{m+u}(\mathcal{V}) = \mathbf{E}_{N_1, N_2} \{s(N_1, N_2)\}$.

Claim 2. $\mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} = s(\mathbf{E}_{\tilde{\sigma}} N_1, \mathbf{E}_{\tilde{\sigma}} N_2)$.

Claim 3. $s(\mathbf{E}_{\tilde{\sigma}} N_1, \mathbf{E}_{\tilde{\sigma}} N_2) - \mathbf{E}_{N_1, N_2} \{s(N_1, N_2)\} \leq c_0 B \left(\frac{1}{u} + \frac{1}{m}\right) \sqrt{m}$.

Having established these three claims we immediately obtain

$$\mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} \leq \tilde{R}_{m+u}(\mathcal{V}) + c_0 B \left(\frac{1}{u} + \frac{1}{m}\right) \sqrt{m}. \quad (8)$$

The entire development is symmetric in m and u and, therefore, we also obtain the same result but with \sqrt{u} instead of \sqrt{m} . By taking the minimum of (8) and the symmetric bound (with \sqrt{u}) we establish the theorem. It remains to prove the three claims.

Proof of Claim 1. Note that N_1 and N_2 are random variables whose distribution is induced by the distribution of $\tilde{\sigma}$. We have

$$\tilde{R}_{m+u}(\mathcal{V}) = \mathbf{E}_{N_1, N_2} \mathbf{E}_{\tilde{\sigma} \sim \text{Rad}(N_1, N_2)} \sup_{\mathbf{v} \in \mathcal{V}} \sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) v(i) = \mathbf{E}_{N_1, N_2} s(N_1, N_2). \quad (9)$$

Proof of Claim 2. (Sketch) By the definitions of \mathbf{H}_k and \mathbf{T}_k (appearing just after Lemma 2), for any $N_1, N_2 \in I_1^{m+u}$ we have

$$\begin{aligned} & \mathbf{E}_{\mathbf{Z}, \mathbf{Z}'} \sup_{\mathbf{v} \in \mathcal{V}} \left[\mathbf{T}_{N_1} \{\mathbf{v}(\mathbf{Z})\} - \mathbf{T}_{N_2} \{\mathbf{v}(\mathbf{Z}')\} + \mathbf{H}_{N_2} \{\mathbf{v}(\mathbf{Z}')\} - \mathbf{H}_{N_1} \{\mathbf{v}(\mathbf{Z})\} \right] = \\ & \mathbf{E}_{\mathbf{Z}, \mathbf{Z}'} \sup_{\mathbf{v} \in \mathcal{V}} \left[\frac{1}{u} \sum_{i=N_1+1}^{m+u} v(Z_i) - \frac{1}{u} \sum_{i=N_2+1}^{m+u} v(Z'_i) + \frac{1}{m} \sum_{i=1}^{N_2} v(Z'_i) - \frac{1}{m} \sum_{i=1}^{N_1} v(Z_i) \right]. \end{aligned} \quad (9)$$

The values of N_1 and N_2 , and the distribution of \mathbf{Z} and \mathbf{Z}' , with respect to which we take the expectation in (9), induce a distribution of assignments of coefficients $\{\frac{1}{m}, -\frac{1}{m}, \frac{1}{u}, -\frac{1}{u}\}$ to the components of \mathbf{v} . For any N_1, N_2 and realizations of \mathbf{Z} and \mathbf{Z}' , each component $v(i)$, $i \in I_1^{m+u}$, is assigned to exactly two coefficients, one for each of the two permutations (\mathbf{Z} and \mathbf{Z}'). Let $\mathbf{a} \triangleq (a_1, \dots, a_{m+u})$, where $a_i \triangleq (a_{i,1}, a_{i,2})$. For any $i \in I_1^{m+u}$, the pair $(a_{i,1}, a_{i,2})$

takes the values of the coefficients of $v(i)$, where the first component is induced by the realization \mathbf{Z} (i.e., $a_{i,1}$ is either $-\frac{1}{m}$ or $\frac{1}{u}$) and the second component by the realization of \mathbf{Z}' (i.e., $a_{i,2}$ is either $\frac{1}{m}$ or $-\frac{1}{u}$).

Let $\mathbf{A}(N_1, N_2)$ be the distribution of vectors \mathbf{a} , induced by the distribution of \mathbf{Z} and \mathbf{Z}' , for particular N_1, N_2 . Using this definition we can write

$$(9) = \mathbf{E}_{\mathbf{a} \sim \mathbf{A}(N_1, N_2)} \sup_{\mathbf{v} \in \mathcal{V}} \left[\sum_{i=1}^{m+u} (a_{i,1} + a_{i,2}) v(i) \right]. \quad (10)$$

We argue (the full proof will appear in the full version) that the distributions $\mathbf{R}(N_1, N_2)$ and $\mathbf{A}(N_1, N_2)$ are identical. Therefore, it follows from (10) that

$$(9) = \mathbf{E}_{\tilde{\sigma} \sim \mathbf{R}(N_1, N_2)} \left\{ \sup_{\mathbf{v} \in \mathcal{V}} \left[\sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) v(i) \right] \right\} = s(N_1, N_2). \quad (11)$$

Note that $\mathbf{E}_{\tilde{\sigma}} N_1 = \mathbf{E}_{\tilde{\sigma}} \{n_1 + n_2\} = m$ and $\mathbf{E}_{\tilde{\sigma}} N_2 = \mathbf{E}_{\tilde{\sigma}} \{n_2 + n_3\} = m$. Hence

$$\mathbf{E}_{\mathbf{Z}} \{g(\mathbf{Z})\} = \mathbf{E}_{\tilde{\sigma} \sim \text{Rad}(m, m)} \left\{ \sup_{\mathbf{v} \in \mathcal{V}} \left[\sum_{i=1}^{m+u} (\tilde{\sigma}_{i,1} + \tilde{\sigma}_{i,2}) v(i) \right] \right\} = s(\mathbf{E}_{\tilde{\sigma}} N_1, \mathbf{E}_{\tilde{\sigma}} N_2).$$

Proof of Claim 3. (Sketch) Abbreviate $Q \triangleq \frac{1}{m} + \frac{1}{u}$. For any $1 \leq N_1, N_2, N'_1, N'_2 \leq m + u$ we have (the technical proof will appear in the full version),

$$|s(N_1, N_2) - s(N'_1, N_2)| \leq B_{\max} |N_1 - N'_1| Q, \quad (12)$$

$$|s(N_1, N_2) - s(N_1, N'_2)| \leq B_{\max} |N_2 - N'_2| Q. \quad (13)$$

We use the following Bernstein-type concentration inequality (see [10], Problem 8.3) for the Binomial random variable $X \sim \text{Bin}(p, n)$: $\mathbf{P}_X \{|X - \mathbf{E}X| > t\} < 2 \exp\left(-\frac{3t^2}{8np}\right)$. Noting that $N_1, N_2 \sim \text{Bin}\left(\frac{m}{m+u}, m+u\right)$, we use (12), (13) and the Bernstein-type inequality (applied with $n \triangleq m+u$ and $p \triangleq \frac{m}{m+u}$) to obtain

$$\begin{aligned} & \mathbf{P}_{N_1, N_2} \{|s(N_1, N_2) - s(\mathbf{E}_{\tilde{\sigma}} \{N_1\}, \mathbf{E}_{\tilde{\sigma}} \{N_2\})| \geq \epsilon\} \\ & \leq \mathbf{P}_{N_1, N_2} \{|s(N_1, N_2) - s(N_1, \mathbf{E}_{\tilde{\sigma}} N_2)| + |s(N_1, \mathbf{E}_{\tilde{\sigma}} N_2) - s(\mathbf{E}_{\tilde{\sigma}} N_1, \mathbf{E}_{\tilde{\sigma}} N_2)| \geq \epsilon\} \\ & \leq \mathbf{P}_{N_1, N_2} \left\{ |s(N_1, N_2) - s(N_1, \mathbf{E}_{\tilde{\sigma}} N_2)| \geq \frac{\epsilon}{2} \right\} \\ & \quad + \mathbf{P}_{N_1, N_2} \left\{ |s(N_1, \mathbf{E}_{\tilde{\sigma}} N_2) - s(\mathbf{E}_{\tilde{\sigma}} N_1, \mathbf{E}_{\tilde{\sigma}} N_2)| \geq \frac{\epsilon}{2} \right\} \\ & \leq \mathbf{P}_{N_2} \left\{ |N_2 - \mathbf{E}_{\tilde{\sigma}} N_2| B_{\max} Q \geq \frac{\epsilon}{2} \right\} + \mathbf{P}_{N_1} \left\{ |N_1 - \mathbf{E}_{\tilde{\sigma}} N_1| B_{\max} Q \geq \frac{\epsilon}{2} \right\} \\ & \leq 4 \exp\left(-\frac{3\epsilon^2}{32(m+u) \frac{m}{m+u} B_{\max}^2 Q^2}\right) = 4 \exp\left(-\frac{3\epsilon^2}{32m B_{\max}^2 Q^2}\right). \end{aligned}$$

Next we use the following fact (see [10], Problem 12.1): if a nonnegative random variable X satisfies $\mathbf{P}\{X > t\} \leq c \cdot \exp(-kt^2)$, then $\mathbf{E}X \leq \sqrt{\ln(ce)/k}$.

Using this fact with $c \triangleq 4$ and $k \triangleq 3/(32mQ^2)$ we have

$$\begin{aligned} |\mathbf{E}_{N_1, N_2} \{s(N_1, N_2)\} - s(\mathbf{E}_{\boldsymbol{\sigma}} N_1, \mathbf{E}_{\boldsymbol{\sigma}} N_2)| &\leq \mathbf{E}_{N_1, N_2} |s(N_1, N_2) - s(\mathbf{E}_{\boldsymbol{\sigma}} N_1, \mathbf{E}_{\boldsymbol{\sigma}} N_2)| \\ &\leq \sqrt{\frac{32 \ln(4e)}{3} m B_{\max}^2 Q^2} . \end{aligned} \quad (14)$$

□

By combining (3) and Lemma 3 we obtain the next concentration inequality, which is the main result of this section.

Theorem 1. *Let $B_1 \leq 0$, $B_2 \geq 0$ and \mathcal{V} be a (possibly infinite) set of real-valued vectors in $[B_1, B_2]^{m+u}$. Let $B \triangleq B_2 - B_1$ and $B_{\max} \triangleq \max(|B_1|, |B_2|)$. Let $Q \triangleq (\frac{1}{u} + \frac{1}{m})$. Then with probability of at least $1 - \delta$ over random permutation \mathbf{Z} of I_1^{m+u} , for all $\mathbf{v} \in \mathcal{V}$,*

$$\mathbf{T}\{\mathbf{v}(\mathbf{Z})\} \leq \mathbf{H}\{\mathbf{v}(\mathbf{Z})\} + R_{m+u}(\mathcal{V}) + B_{\max} c_0 Q \sqrt{\min(m, u)} + B \sqrt{2Q \ln \frac{1}{\delta}}. \quad (15)$$

4 Uniform Rademacher error bound

Our goal now is to utilize the concentration inequality of Theorem 1 to derive a uniform error bound for all soft labelings $\mathbf{h} \in \mathcal{H}$ of the full-sample. The idea is to apply Theorem 1 with an appropriate instantiation of the set \mathcal{V} so that $\mathbf{T}\{\mathbf{v}(\mathbf{Z})\}$ will correspond to the test error and $\mathbf{H}\{\mathbf{v}(\mathbf{Z})\}$ to the empirical error. The following lemma will be used in this analysis. The lemma is an adaptation, which accommodates the transductive Rademacher variables, of Lemma 5 from [17]. The proof is omitted (but will be provided in the full version).

Lemma 4. *Let $\mathcal{H} \subseteq \mathbb{R}^{m+u}$ be a set of vectors. Let f and g be real-valued functions. Let $\boldsymbol{\sigma} = \{\sigma_i\}_{i=1}^{m+u}$ be Rademacher variables, as defined in (1). If for all $1 \leq i \leq m+u$ and any $\mathbf{h}, \mathbf{h}' \in \mathcal{H}$, $|f(h(i)) - f(h'(i))| \leq |g(h(i)) - g(h'(i))|$, then*

$$\mathbf{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{h} \in \mathcal{H}} \left[\sum_{i=1}^{m+u} \sigma_i f(h(i)) \right] \leq \mathbf{E}_{\boldsymbol{\sigma}} \sup_{\mathbf{h} \in \mathcal{H}} \left[\sum_{i=1}^{m+u} \sigma_i g(h(i)) \right]. \quad (16)$$

Let $Y \in \{\pm 1\}^{m+u}$, and denote by $Y(i)$ the i th component of Y . For any Y define $\ell_\gamma^Y(h(i)) \triangleq \ell_\gamma(h(i), Y(i))$. Noting that ℓ_γ^Y satisfies the Lipschitz condition $|\ell_\gamma^Y(h(i)) - \ell_\gamma^Y(h'(i))| \leq \frac{1}{\gamma} |h(i) - h'(i)|$, we apply Lemma 4 with the functions $f(h(i)) = \ell_\gamma^Y(h(i))$ and $g(h(i)) = h(i)/\gamma$, to get

$$\mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\mathbf{h} \in \mathcal{H}} \sum_{i=1}^{m+u} \sigma_i \ell_\gamma^Y(h(i)) \right\} \leq \frac{1}{\gamma} \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\mathbf{h} \in \mathcal{H}} \sum_{i=1}^{m+u} \sigma_i h(i) \right\}. \quad (17)$$

For any Y , define $\boldsymbol{\ell}_\gamma^Y(\mathbf{h}) \triangleq (\ell_\gamma^Y(h(1)), \dots, \ell_\gamma^Y(h(m+u)))$. Taking Y to be the true (unknown) labeling of the full-sample, we set $L_{\mathcal{H}}^\gamma = \{\mathbf{v} : \mathbf{v} = \boldsymbol{\ell}_\gamma^Y(\mathbf{h}), \mathbf{h} \in \mathcal{H}\}$.

It follows from (17) that $R_{m+u}(L_{\mathcal{H}}^{\gamma}) \leq \frac{1}{\gamma}R_{m+u}(\mathcal{H})$. Applying Theorem 1 with $\mathbf{v} \triangleq \ell_{\gamma}(\mathbf{h})$, $\mathcal{V} \triangleq L_{\mathcal{H}}^{\gamma}$, $B_{\max} = B = 1$, and using the last inequality we obtain:²

Theorem 2. *Let \mathcal{H} be any set of full-sample soft labelings. The choice of \mathcal{H} can depend on the full-sample X_{m+u} . Let $c_0 = \sqrt{\frac{32 \ln(4e)}{3}} < 5.05$ and $Q \triangleq (\frac{1}{u} + \frac{1}{m})$. For any fixed γ , with probability of at least $1 - \delta$ over the choice of the training set from X_{m+u} , for all $\mathbf{h} \in \mathcal{H}$,*

$$\mathcal{L}_u(\mathbf{h}) \leq \mathcal{L}_u^{\gamma}(\mathbf{h}) \leq \widehat{\mathcal{L}}_m^{\gamma}(\mathbf{h}) + \frac{R_{m+u}(\mathcal{H})}{\gamma} + c_0 Q \sqrt{\min(m, u)} + \sqrt{2Q \ln \frac{1}{\delta}}. \quad (18)$$

5 Bounds for Unlabeled-Labeled Decompositions (ULDs)

Let r be any natural number and let K be an $(m+u) \times r$ matrix depending only on X_{m+u} . Let $\boldsymbol{\alpha}$ be an $r \times 1$ vector that may depend on both S_m and X_u . The soft classification output \mathbf{y} of any transductive algorithm can be represented by

$$\mathbf{y} = K \cdot \boldsymbol{\alpha}. \quad (19)$$

We refer to (19) as an *unlabeled-labeled decomposition (ULD)*. In this section we develop bounds on the Rademacher complexity of algorithms based on their ULDs. We note that any transductive algorithm has a trivial ULD, for example, by taking $r = m + u$, setting K to be the identity matrix and assigning $\boldsymbol{\alpha}$ to any desired (soft) labels. We are interested in “non-trivial” ULDs and provide useful bounds for such decompositions.³

In a “vanilla” ULD, K is an $(m+u) \times (m+u)$ matrix and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_{m+u})$ simply specifies the given labels in S_m (where $\alpha_i \in \{\pm 1\}$ for labeled points, and $\alpha_i = 0$ otherwise). From our point of view any vanilla ULD is not trivial because $\boldsymbol{\alpha}$ does not encode the final classification of the algorithm. For example, the algorithm of Zhou et al. [22] straightforwardly admits a vanilla ULD. On the other hand, the natural (non-trivial) ULD of the algorithms of Zhu et al. [23] and of Belkin and Niyogi [5] is not of the vanilla type. For some algorithms it is not necessarily obvious how to find non-trivial ULDs. Later we mention such cases – in particular, the algorithms of Joachims [15] and of Belkin et al. [4].

We now present a bound on the transductive Rademacher complexity of any transductive algorithm basing on their ULD. Let $\{\lambda_i\}_{i=1}^r$ be the singular values of K . We use the well-known fact that $\|K\|_{\text{Fro}} = \sqrt{\sum_{i=1}^r \lambda_i^2}$, where $\|K\|_{\text{Fro}} \triangleq \sqrt{\sum_{i,j} (K(i,j))^2}$ is the Frobenius norm of K . Suppose that $\|\boldsymbol{\alpha}\|_2 \leq \mu_1$ for some μ_1 . Let $\mathcal{H} \triangleq \mathcal{H}(K)$ be the transductive hypothesis space induced by the matrix

² This bound holds for any *fixed* margin parameter γ . Using the technique of the proof of Theorem 18 in [7] we can also obtain a bound that is uniform in γ .

³ For the trivial decomposition where K is the identity matrix it can be shown that the risk bound (18), combined with the forthcoming Rademacher complexity bound (20), is greater than 1 (the proof will be provided in the full version).

K ; that is, \mathcal{H} is the set of all possible outputs of the algorithm corresponding to a fixed full-sample X_{m+u} , all possible training/test partitions and all possible labelings of the training set. Using the abbreviation $K(i, \cdot)$ for the i th row of K and following the proof idea of Lemma 22 in [3], we obtain (the complete derivation will appear in the full version),

$$\begin{aligned} R_{m+u}(\mathcal{H}) &= \mathbf{E}_{\sigma} \left\{ \sup_{h \in \mathcal{H}} \sum_{i=1}^{m+u} \sigma_i h(x_i) \right\} = \mathbf{E}_{\sigma} \left\{ \sup_{\alpha: \|\alpha\|_2 \leq \mu_1} \sum_{i=1}^{m+u} \sigma_i \langle \alpha, K(i, \cdot) \rangle \right\} \\ &\leq \mu_1 \sqrt{\sum_{i=1}^{m+u} \frac{2}{mu} \langle K(i, \cdot), K(i, \cdot) \rangle} = \mu_1 \sqrt{\frac{2}{mu} \|K\|_{\text{Fro}}^2} = \mu_1 \sqrt{\frac{2}{mu} \sum_{i=1}^r \lambda_i^2}, \quad (20) \end{aligned}$$

where the inequality is obtained using the Cauchy-Schwartz and Jensen inequalities. Using the bound (20) in conjunction with Theorem 2 we get a data-dependent error bound for any algorithm, that can be computed once we derive an upper bound on the maximal length of possible values of the α vector, appearing in its ULD. Notice that for any vanilla ULD, $\mu_1 = \sqrt{m}$. Later on we derive a tight μ_1 for non-trivial ULDs of SGT [15] and of the ‘‘consistency method’’ [22].

The bound (20) is syntactically similar in form to a corresponding inductive Rademacher bound of kernel machines [3]. However, as noted above, the fundamental difference is that in induction, the choice of the kernel (and therefore \mathcal{H}) must be *data-independent* in the sense that it must be selected before the training examples are observed. In our transductive setting, K and \mathcal{H} can be selected based on the unlabeled full-sample.

5.1 Example: Analysis of SGT

We now exemplify the use of the ULD Rademacher bound (20) and analyze the SGT algorithm [15]. We start with a description of a simplified version of SGT that captures the essence of the algorithm.⁴ Let W be a symmetric $(m+u) \times (m+u)$ similarity matrix of the full-sample X_{m+u} . The matrix W can be built in various ways, for example, it can be a k -nearest neighbors graph. Let D be a diagonal matrix, whose (i, i) th entry is the sum of the i th row in W . An unnormalized Laplacian of W is $L = D - W$. Let $\tau = (\tau_1, \dots, \tau_{m+u})$ be a vector that specifies the given labels in S_m ; that is, $\tau_i \in \{\pm 1\}$ for labeled points, and $\tau_i = 0$ otherwise. Let c be a fixed constant and $\mathbf{1}$ be an $(m+u) \times 1$ vector whose entries are 1 and let C be a diagonal matrix such that $C(i, i) = 1$ iff example i is in the training set (and zero otherwise). The soft classification \mathbf{h}^* produced by the SGT algorithm is the solution of the following optimization problem:

$$\min_{\mathbf{h} \in \mathbb{R}^{m+u}} \mathbf{h}^T L \mathbf{h} + c(\mathbf{h} - \tau)^T C(\mathbf{h} - \tau) \quad (21)$$

$$s.t. \mathbf{h}^T \mathbf{1} = 0, \quad \mathbf{h}^T \mathbf{h} = m + u. \quad (22)$$

⁴ We omit some heuristics that are optional in SGT. Their inclusion does not affect the error bound we derive.

It is shown in [15] that $\mathbf{h}^* = K\boldsymbol{\alpha}$, where K is an $(m+u) \times r$ matrix⁵ whose columns are orthonormal eigenvectors corresponding to non-zero eigenvalues of the Laplacian L and $\boldsymbol{\alpha}$ is an $r \times 1$ vector. While $\boldsymbol{\alpha}$ depends on both the training and test sets, the matrix K depends only on the unlabeled full-sample. Substituting $\mathbf{h}^* = K\boldsymbol{\alpha}$ to the second constraint in (22) and using the orthonormality of the columns of K , we get $m+u = \mathbf{h}^T \mathbf{h} = \boldsymbol{\alpha}^T K^T K \boldsymbol{\alpha} = \boldsymbol{\alpha}^T \boldsymbol{\alpha}$. Hence, $\|\boldsymbol{\alpha}\|_2 = \sqrt{m+u}$ and we can take $\mu_1 = \sqrt{m+u}$. Since K is an $(m+u) \times r$ matrix with orthonormal columns, $\|K\|_{\text{Fro}}^2 = r$. Consequently, by (20) the transductive Rademacher complexity of SGT is upper bounded by $\sqrt{2r \left(\frac{1}{m} + \frac{1}{u}\right)}$, where r is the number of non-zero eigenvalues of L . Notice that this bound is oblivious to the magnitude of these eigenvalues.

5.2 Kernel ULD

If $r = m+u$ and K is a kernel matrix (this holds if K is positive semidefinite), then we say that the decomposition is a *kernel-ULD*. Let $\mathcal{H} \subseteq \mathbb{R}^{m+u}$ be the reproducing kernel Hilbert space (RKHS), corresponding to K . We denote by $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ the inner product in \mathcal{H} . Since K is a kernel matrix, by the reproducing property⁶ of \mathcal{H} , $K(i, j) = \langle K(i, \cdot), K(j, \cdot) \rangle_{\mathcal{H}}$. Suppose that the vector $\boldsymbol{\alpha}$ satisfies $\sqrt{\boldsymbol{\alpha}^T K \boldsymbol{\alpha}} \leq \mu_2$ for some μ_2 . Let $\{\lambda_i\}_{i=1}^{m+u}$ be the eigenvalues of K . By similar arguments used to derive (20) we have (details will appear in the full version):

$$\begin{aligned} R_{m+u}(\mathcal{H}) &= \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{h \in \mathcal{H}} \sum_{i=1}^{m+u} \sigma_i h(x_i) \right\} = \mathbf{E}_{\boldsymbol{\sigma}} \left\{ \sup_{\boldsymbol{\alpha}: \sqrt{\boldsymbol{\alpha}^T K \boldsymbol{\alpha}} \leq \mu_2} \sum_{i,j=1}^{m+u} \sigma_i \alpha_j K(i, j) \right\} \\ &\leq \mu_2 \sqrt{\sum_{i=1}^{m+u} \frac{2}{mu} K(i, i)} = \mu_2 \sqrt{\frac{2 \cdot \text{trace}(K)}{mu}} = \mu_2 \sqrt{\frac{2}{mu} \sum_{i=1}^{m+u} \lambda_i}. \end{aligned} \quad (23)$$

By defining the RKHS induced by the unnormalized Laplacian, as in [14], and using a generalized representer theorem [18], it can be shown that the algorithm of Belkin et al. [4] has a kernel-ULD (the details will appear in the full version).

5.3 Monte-Carlo Rademacher bounds

We now show how to compute Monte-Carlo Rademacher bounds with high confidence for any transductive algorithm using its ULD. Our empirical examination of these bounds shows that they are tighter than the analytical bounds (20) and (23). The technique, which is based on a simple application of Hoeffding's inequality, is made particularly simple for vanilla ULDs.

Let $\mathcal{V} \subseteq \mathbb{R}^{m+u}$ be a set of vectors, $\boldsymbol{\sigma} \in \mathbb{R}^{m+u}$ to be a Rademacher vector (1), and $g(\boldsymbol{\sigma}) = \sup_{\mathbf{v} \in \mathcal{V}} \boldsymbol{\sigma} \cdot \mathbf{v}$. By Def. 1, $R_{m+u}(\mathcal{V}) = \mathbf{E}_{\boldsymbol{\sigma}} \{g(\boldsymbol{\sigma})\}$. Let $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n$ be an

⁵ r is the number of non-zero eigenvalues of L , after performing spectral transformations. Joachims set the default r to 80.

⁶ This means that $\forall \mathbf{h} \in \mathcal{H}$ and $i \in I_1^{m+u}$, $h(i) = \langle K(i, \cdot), \mathbf{h} \rangle_{\mathcal{H}}$.

i.i.d. sample of Rademacher vectors. We estimate $R_{m+u}(\mathcal{V})$ with high-confidence by applying the Hoeffding inequality on $\sum_{i=1}^n \frac{1}{n} g(\boldsymbol{\sigma}_i)$. To apply the Hoeffding inequality we need a bound on $\sup_{\boldsymbol{\sigma}} |g(\boldsymbol{\sigma})|$, which is derived for the case where \mathcal{V} is all possible outputs of the algorithm (for a fixed X_{m+u}). Specifically, suppose that $\mathbf{v} \in \mathcal{V}$ is an output of the algorithm, $\mathbf{v} = K\boldsymbol{\alpha}$, and assume that $\|\boldsymbol{\alpha}\|_2 \leq \mu_1$. By Def. 1, for all $\boldsymbol{\sigma}$, $\|\boldsymbol{\sigma}\|_2 \leq b \triangleq \sqrt{m+u} \left(\frac{1}{m} + \frac{1}{u}\right)$. Using elementary linear algebra we have $\sup_{\boldsymbol{\sigma}} |g(\boldsymbol{\sigma})| \leq \sup_{\|\boldsymbol{\sigma}\|_2 \leq b, \|\boldsymbol{\alpha}\|_2 \leq \mu_1} |\boldsymbol{\sigma} K \boldsymbol{\alpha}| \leq b \mu_1 \lambda_{\max}$, where λ_{\max} is a maximal singular value of K . Applying the one-sided Hoeffding inequality on n samples of $g(\boldsymbol{\sigma})$ we have, for any given δ , that with probability of at least $1 - \delta$ over the random i.i.d. choice of the vectors $\boldsymbol{\sigma}_1, \dots, \boldsymbol{\sigma}_n$,

$$R_{m+u}(\mathcal{V}) \leq \frac{1}{n} \sum_{i=1}^n \sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 \leq \mu_1} \boldsymbol{\sigma}_i K \boldsymbol{\alpha} + \mu_1 \sqrt{m+u} \left(\frac{1}{m} + \frac{1}{u}\right) \lambda_{\max} \sqrt{\frac{2 \ln \frac{1}{\delta}}{n}}. \quad (24)$$

To use the bound (24), the value of $\sup_{\boldsymbol{\alpha}: \|\boldsymbol{\alpha}\|_2 \leq \mu} \boldsymbol{\sigma}_i K \boldsymbol{\alpha}$ should be computed for each randomly drawn $\boldsymbol{\sigma}_i$. This computation is algorithm-dependent and below we show how to compute it for the algorithm of [22].⁷ In cases where we can compute the supremum exactly (as in vanilla ULDs; see below) we can also get a lower bound using the symmetric Hoeffding inequality.

Example: Application to the CM algorithm. We start with a brief description of the Consistency Method (CM) algorithm of [22]. The algorithm has a natural vanilla ULD (see definition at the beginning of Sec. 5), where the matrix K is computed as follows. Let W and D be matrices as in SGT (see Sec. 5.1). A normalized Laplacian of W is $L = D^{-1/2} W D^{-1/2}$. Let β be a parameter in $(0, 1)$. Then, $K \triangleq (1 - \beta)(I - \beta L)^{-1}$ and the output of CM is $\mathbf{y} = K \cdot \boldsymbol{\alpha}$, where $\boldsymbol{\alpha}$ specifies the given labels. Consequently $\|\boldsymbol{\alpha}\|_2 \leq \sqrt{m}$. Moreover, it can be verified that K is a kernel matrix, and therefore, the decomposition is a kernel-ULD. It turns out that for CM, the exact value of the supremum in (24) can be analytically derived. The vectors $\boldsymbol{\alpha}$, that induce the CM hypothesis space for a particular K , have exactly m components with values in $\{\pm 1\}$; the rest of the components are zeros. Let Ψ be the set of all possible such $\boldsymbol{\alpha}$'s. Let $\mathbf{t}(\boldsymbol{\sigma}_i) = (t_1, \dots, t_{m+u}) \triangleq \boldsymbol{\sigma}_i K \in \mathbb{R}^{1 \times (m+u)}$ and $|\mathbf{t}(\boldsymbol{\sigma}_i)| \triangleq (|t_1|, \dots, |t_{m+u}|)$. Then, for any fixed $\boldsymbol{\sigma}_i$, $\sup_{\boldsymbol{\alpha} \in \Psi} \boldsymbol{\sigma}_i K \boldsymbol{\alpha}$ is the sum of the m largest elements in $|\mathbf{t}(\boldsymbol{\sigma}_i)|$. This derivation holds for any vanilla ULD.

To demonstrate the Rademacher bounds discussed in this paper we present an empirical comparison of the bounds over two datasets (**Voting** and **Pima**) from the UCI repository. For each dataset we took $m+u$ to be the size of the dataset (435 and 768, respectively) and we took m to be 1/3 of the full-sample size. The matrix W is the 10-nearest neighbor graph computed with the cosine similarity metric. We applied the CM algorithm with $\beta = 0.5$. The Monte-Carlo bounds (both upper and lower) were computed with $\delta = 0.05$ and $n = 10^5$.

⁷ An application of this approach in induction seems to be very hard, if not impossible. For example, in the case of RBF kernel machines we will need to optimize over (typically) infinite-dimensional vectors in the feature space.

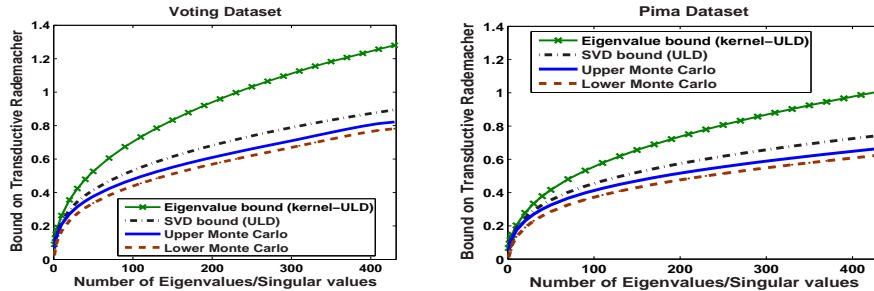


Fig. 1. A comparison of transductive Rademacher bounds.

We compared the Monte-Carlo bounds with the ULD bound (20), named here “the SVD bound”, and the kernel-ULD bound (23), named here “the eigenvalue bound”. The graphs in Figure 1 compare these four bounds for each of the datasets as a function of the number of non-zero eigenvalues of K (trimmed to maximum 430 eigenvalues). Specifically, each point t on the x -axis corresponds to bounds computed with a matrix K_t that approximates K using only the smallest t eigenvalues of K . In both examples the lower and upper Monte-Carlo bounds tightly “sandwich” the true Rademacher complexity. It is striking that the SVD bound is very close to the true Rademacher complexity. In principle, with our simple Monte-Carlo method we can approximate the true Rademacher complexity up to any desired accuracy (with high confidence) at the cost of drawing sufficiently many Rademacher vectors.

6 PAC-Bayesian bound for transductive mixtures

In this section we adapt part of the results of [17] to transduction. The proofs of all results presented in this section will appear in the full version of the paper.

Let $\mathcal{B} = \{\mathbf{h}_i\}_{i=1}^{|\mathcal{B}|}$ be a finite set of *base-hypotheses*. The class \mathcal{B} can be formed after observing the full-sample X_{m+u} , but before obtaining the training/test set partition and the labels. Let $\mathbf{q} = (q_1, \dots, q_{|\mathcal{B}|}) \in \mathbb{R}^{|\mathcal{B}|}$. Our goal is to construct a useful *mixture hypothesis*, $\tilde{\mathbf{h}}_{\mathbf{q}} \triangleq \sum_{i=1}^{|\mathcal{B}|} q_i \mathbf{h}_i$. We assume that \mathbf{q} belongs to a domain $\Omega_{g,A} = \{\mathbf{q} \mid g(\mathbf{q}) \leq A\}$, where $g: \mathbb{R}^{|\mathcal{B}|} \rightarrow \mathbb{R}$ is a predefined function and $A \in \mathbb{R}$ is a constant. The domain $\Omega_{g,A}$ and the set \mathcal{B} induce the class $\tilde{\mathcal{B}}_{g,A}$ of all possible mixtures $\tilde{\mathbf{h}}_{\mathbf{q}}$. Recalling that $Q \triangleq (1/m + 1/u)$ and $c_0 = \sqrt{32 \ln(4e)}/3 < 5.05$, we apply Theorem 2 with $\mathcal{H} \triangleq \tilde{\mathcal{B}}_{g,A}$ and obtain that with probability of at least $1 - \delta$ over the training/test partition of X_{m+u} , for all $\tilde{\mathbf{h}}_{\mathbf{q}} \in \tilde{\mathcal{B}}_{g,A}$,

$$\mathcal{L}_u(\tilde{\mathbf{h}}_{\mathbf{q}}) \leq \hat{\mathcal{L}}_m^\gamma(\tilde{\mathbf{h}}_{\mathbf{q}}) + \frac{R_{m+u}(\tilde{\mathcal{B}}_{g,A})}{\gamma} + c_0 Q \sqrt{\min(m, u)} + \sqrt{2Q \ln \frac{1}{\delta}}. \quad (25)$$

Let $Q_1 \triangleq \sqrt{2Q (\ln(1/\delta) + 2 \ln \log_s (s\hat{g}(\mathbf{q})/g_0))}$. It is straightforward to apply the technique used in the proof of Theorem 10 in [17] and obtain the following bound, which eliminates the dependence on A .

Corollary 1. *Let $g_0 > 0$, $s > 1$ and $\tilde{g}(\mathbf{q}) = s \max(g(\mathbf{q}), g_0)$. For any (fixed) g , with probability of at least $1 - \delta$ over the training/test set partition, for all⁸ $\tilde{\mathbf{h}}_{\mathbf{q}}$,*

$$\mathcal{L}_u(\tilde{\mathbf{h}}_{\mathbf{q}}) \leq \hat{\mathcal{L}}_m^\gamma(\tilde{\mathbf{h}}_{\mathbf{q}}) + \frac{R_{m+u}(\tilde{\mathcal{B}}_{g, \tilde{g}(\mathbf{q})})}{\gamma} + c_0 Q \sqrt{\min(m, u)} + Q_1 . \quad (26)$$

We now instantiate Corollary 1 for $g(\mathbf{q})$ being the KL-divergence and derive a PAC-Bayesian bound. To this end, we restrict \mathbf{q} to be a probability vector. Let $\mathbf{p} \in \mathbb{R}^{|\mathcal{B}|}$ be a “prior” probability vector. The vector \mathbf{p} can only depend on the unlabeled full-sample X_{m+u} . For a particular prior \mathbf{p} let $g(\mathbf{q}) \triangleq D(\mathbf{q} \parallel \mathbf{p}) = \sum_{i=1}^{|\mathcal{B}|} q_i \ln \left(\frac{q_i}{p_i} \right)$. Adopting Lemma 11 of [17] to the transductive Rademacher variables, defined in (1), we obtain the following bound.

Theorem 3. *Let $g_0 > 0$, $s > 1$. Let \mathbf{p} and \mathbf{q} be any prior and posterior distribution over \mathcal{B} , respectively. Set $g(\mathbf{q}) \triangleq D(\mathbf{q} \parallel \mathbf{p})$ and $\tilde{g}(\mathbf{q}) \triangleq s \max(g(\mathbf{q}), g_0)$. Then, with prob. of at least $1 - \delta$ over the training/test set partition, for all $\tilde{\mathbf{h}}_{\mathbf{q}}$,*

$$\mathcal{L}_u(\tilde{\mathbf{h}}_{\mathbf{q}}) \leq \hat{\mathcal{L}}_m^\gamma(\tilde{\mathbf{h}}_{\mathbf{q}}) + \frac{Q}{\gamma} \sqrt{2\tilde{g}(\mathbf{q}) \sup_{\mathbf{h} \in \mathcal{B}} \|\mathbf{h}\|_2^2} + c_0 Q \sqrt{\min(m, u)} + Q_1 . \quad (27)$$

Theorem 3 is a PAC-Bayesian result, where the prior \mathbf{p} can depend on X_{m+u} and the posterior can be optimized adaptively, based also on S_m .

7 Concluding remarks

We have studied the use of Rademacher complexity analysis in the transductive setting. Our results include the first general Rademacher bound for soft classification algorithms, the unlabeled-labeled decomposition (ULD) technique for bounding Rademacher complexity of any transductive algorithm and a bound for Bayesian mixtures.

It would be nice to further improve our bounds using, for example, the local Rademacher approach [2]. However, we believe that the main advantage of these transductive bounds is the possibility of selecting a hypothesis space based on the full-sample. A clever data-dependent choice of this space should provide sufficient flexibility to achieve a low training error with low Rademacher complexity. In our opinion this opportunity can be explored and exploited much further.

This work opens up new avenues for future research. For example, it would be interesting to optimize the matrix K in the ULD representation explicitly (to fit the data) under a constraint of low Rademacher complexity. Also, it would be nice to find “low-Rademacher” approximations of particular K matrices. The PAC-Bayesian bound for mixture algorithms motivates the development and use of transductive mixtures, an area that has yet to be investigated.

Acknowledgement. We thank Yair Wiener for useful comments.

⁸ In the bound (26) the meaning of $R_{m+u}(\tilde{\mathcal{B}}_{g, \tilde{g}(\mathbf{q})})$ is as follows. For any \mathbf{q} let $A = \tilde{g}(\mathbf{q})$ and $R_{m+u}(\tilde{\mathcal{B}}_{g, \tilde{g}(\mathbf{q})}) \triangleq R_{m+u}(\tilde{\mathcal{B}}_{g, A})$.

References

1. M.F. Balcan and A. Blum. An augmented pac model for semi-supervised learning. In O. Chapelle, B. Schölkopf, and A. Zien, editors, *Semi-Supervised Learning*, chapter 22, pages 383–404. MIT Press, 2006.
2. P. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Probability*, 33(4):1497–1537, 2005.
3. P. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.
4. M. Belkin, I. Matveeva, and P. Niyogi. Regularization and semi-supervised learning on large graphs. In *COLT*, pages 624–638, 2004.
5. M. Belkin and P. Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56:209–239, 2004.
6. A. Blum and J. Langford. PAC-MDL Bounds. In *COLT*, pages 344–357, 2003.
7. O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, 2002.
8. O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.
9. P. Derbeko, R. El-Yaniv, and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.
10. L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.
11. R. El-Yaniv and L. Gerzon. Effective transductive learning via objective model selection. *Pattern Recognition Letters*, 26:2104–2115, 2005.
12. R. El-Yaniv and D. Pechyony. Stable transductive learning. In G. Lugosi and H.U. Simon, editors, *Proceedings of the 19th Annual Conference on Learning Theory*, pages 35–49, 2006.
13. S. Hanneke. An analysis of graph cut size for transductive learning. In *ICML*, pages 393–399, 2006.
14. M. Herbster, M. Pontil, and L. Wainer. Online learning over graphs. In *ICML*, pages 305–312, 2005.
15. T. Joachims. Transductive learning via spectral graph partitioning. In *Proceedings of the 20th International Conference on Machine Learning*, pages 290–297, 2003.
16. G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.
17. R. Meir and T. Zhang. Generalization error bounds for Bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.
18. B. Schölkopf, R. Herbrich, and A. Smola. A generalized representer theorem. In D. Helmbold and B. Williamson, editors, *COLT/EuroCOLT*, pages 416–426, 2001.
19. V. Vapnik and A. Chervonenkis. *The theory of pattern recognition*. Moscow: Nauka, 1974.
20. V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, 1982.
21. T. Zhang and R. Ando. Analysis of spectral kernel design based semi-supervised learning. In *NIPS*, pages 1601–1608, 2005.
22. D. Zhou, O. Bousquet, T.N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328, 2003.
23. X. Zhu, Z. Ghahramani, and J.D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.