# Transductive Rademacher Complexities for Learning over a Graph

**K. Pelckmans, J.A.K. Suykens**
K.U.Leuven - ESAT - SCD/SISTA
Kasteelpark Arenberg 10, B-3001, Leuven (Heverlee), Belgium
kristiaan.pelckmans@esat.kuleuven.be

## Abstract

Recent investigations [12, 2, 8, 5, 6] and [11, 9] indicate the use of a probabilistic ('learning') perspective of tasks defined on a single graph, as opposed to the traditional algorithmical ('computational') point of view. This note discusses the use of Rademacher complexities in this setting, and illustrates the use of Kruskal's algorithm for transductive inference based on a nearest neighbor rule.

## 1 Introduction

Weighted, undirected graphs are a widely applicable means for representing domain knowledge of tasks which are defined over finite universes. This manuscript concerns the following question: *'given a fixed graph with (hidden) labels, and a class of plausible labeling over all nodes, if an algorithm is presented with a random subset of labels (of fixed size), how well would the algorithm's* learned *hypothesis contained in this class perform on the remaining labels?'* This question is sometimes termed *distribution-free* transductive setting [12]. The important deviation from the classical inductive setting is that the finite nodes where the hypothesis are to be evaluated are known a-priori. The precise setting also deviates from a *distributional* transductive inference setting where the algorithm has no prior knowledge of the relevant graph, or equivalently, of the inputs. Both settings are discussed and related in [12], and more recently in [3]. Note that the class of plausible labelings (the *hypothesis class*) is not required to contain the 'true' labels (agnostic case).

The following example in collaborative filtering illustrates the practical importance of this problem. Consider a finite set of products which are after careful study organized in an appropriate undirected weighted graph (e.g. based on similarities between products). Let customer A indicate his (binary) preference for a number of random products, and let our algorithm try to fill in his preferences over the remaining products. Subsequently, let customer B do the same and use the algorithm to fill in his unexpressed preferences. After iteration of this scheme for customers A,B,C,D,..., the study of transductive inference characterizes what can be said on the average performance of this scheme.

Some notation is introduced. Let a weighted undirected graph $\mathcal{G}_n = (V, E)$ consist of $1 < n < \infty$ nodes $V = \{v_i\}_{i=1}^n$ with edges $E = \{x_{ij}\}_{i \neq j}$ with $x_{ij} \geq 0$ connected to $v_i$ and $v_j$ for any $i \neq j = 1, \ldots, n$. Assume that no loops occur in the graph, i.e. $x_{ii} = 0$ for all $i = 1, \ldots, n$, and that the graph $\mathcal{G}$ is connected, i.e. there exists a path between any two nodes (this is for notational convenience, most results are valid beyond this restriction). Let $X \in \mathbb{R}^{n \times n}$ denote the positive symmetric matrix defined as $X_{ij} = X_{ji} = x_{ij}$ for all $i, j = 1, \ldots, n$. The Laplacian of $\mathcal{G}$ is then defined as $L = \mathrm{diag}(X 1_n) - X \in \mathbb{R}^{n \times n}$. This paper considers problems where each node has a fixed corresponding label $y_i \in \{-1, 1\}$ such that $\{(v_i, y_i)\}_{i=1}^n$ (or shortly $\{y_i\}_{i=1}^n$), but only a subset $\mathcal{S}_m \subseteq \{1, \ldots, n\}$ with $|\mathcal{S}_m| = m$ of the labels is observed. The task in transductive inference is to predict the labels of the unlabeled nodes $\mathcal{S}_{-m} = \{1, \ldots, n\} \backslash \mathcal{S}_m$. This paper uses the notation $q \in \{-1, 1\}^n$ to denote a hypothesis $\{(v_i, q_i)\}_{i=1}^n$ (or equivalently $\{q_i\}_{i=1}^n$) of the true labeling

$y \in \{-1,1\}^n$ (or $\{y_i\}_{i=1}^n$). The following generic class of hypothesis is studied

$$\mathcal{H} \subseteq \{q \in \{-1,1\}^n\}, \tag{1}$$

with cardinality $|\mathcal{H}|$. It becomes clear that the cardinality of such a class without further specification is $2^n$, which is clearly much too large for reasonable applications. The following sections discusses how one can respectively characterize and construct an appropriate subset for the purpose of learning.

## 2 Transductive Rademacher Complexities

Given a fixed hypothesis set $\mathcal{H}$ of an observed graph $\mathcal{G}$. The risk of an hypothesis $q \in \mathcal{H}$ can be defined here as

$$\mathcal{R}(q|\mathcal{G}) = E_L\left[y_L q_L \middle| \mathcal{G}\right] = \frac{1}{n}\sum_{i=1}^{n} y_i q_i, \tag{2}$$

where the expectation $E_L$ concerns the (uniformly) random index $L \in \{1,\ldots,n\}$. The empirical counterpart becomes $\mathcal{R}_{\mathcal{S}_m}(q|\mathcal{G}) = \frac{1}{m}\sum_{i\in\mathcal{S}_m} y_i q_i$ where $m = |\mathcal{S}_m|$. Several authors discuss generalization bounds suited for the transductive setting [12, 2, 8, 5, 6]. Serfling's inequality provides a convenient way to derive a probabilistic guarantee on the result, see e.g. [8, 9, 10]. Also, Rademacher complexities can be used to give a generalization bound on the result of transductive inference.

**Definition 1 (Rademacher complexity of $\mathcal{H}$ and $\mathcal{G}$)** *Assume that $\mathcal{H}$ is symmetric, i.e. $-\mathcal{H} = \mathcal{H}$. Let $\{\sigma_i\}_{i=1}^n$ be independent random Rademacher variables with $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.*

$$R(\mathcal{H}|\mathcal{G}) = E\left[\sup_{q\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i q_i \;\middle|\; \mathcal{G}\right]. \tag{3}$$

Remark that this definition is more in line to the definition of the inductive case, as opposed to [5] where $R(\mathcal{H})$ is expressed in terms of the random variables $\sigma \in \{-1,0,1\}$. The generalization error of a $q \in \mathcal{H}$ can then be bound as follows

**Theorem 1 (Transductive Rademacher Bound)** *With probability exceeding $1-\delta < 1$, one has for all $q \in \mathcal{H}$ simultaneously*

$$\mathcal{R}_{\mathcal{S}_{-m}}(q|\mathcal{G}) \leq \mathcal{R}_{\mathcal{S}_m}(q|\mathcal{G}) + 2\left(\frac{n}{n-m}\right) R(\mathcal{H}|\mathcal{G}) + 2\sqrt{\left(\frac{1}{m} + \frac{1}{n-m}\right)\log(1/\delta)} \tag{4}$$

The proof goes along the same lines as the classical expression given in [1], but improves on using the martingale expression as in [5] instead of McDiarmids, and secondly by interpreting the symmetrization argument for this transductive setting.

*Proof:* At first, consider the random variable

$$D_{\mathcal{S}_m} = \sup_{q\in\mathcal{H}} \left(\mathcal{R}_{\mathcal{S}_{-m}}(q|\mathcal{G}) - \mathcal{R}_{\mathcal{S}_m}(q|\mathcal{G})\right) \tag{5}$$

and the related quantity $D^{il}_{\mathcal{S}_m} = \sup_{q\in\mathcal{H}}\left(\mathcal{R}_{\mathcal{S}_{-m}{}^{li}}(q|\mathcal{G}) - \mathcal{R}_{\mathcal{S}_m{}^{il}}(q|\mathcal{G})\right)$ where $\mathcal{S}_m{}^{il}$ is defined as $\mathcal{S}_m{}^{il} = \mathcal{S}_m/i \cup l$ where $i \in \mathcal{S}_m$ and $l \in \mathcal{S}_{-m}$. Now it becomes clear that $\beta \triangleq \sup_{i,l}\left|D_{\mathcal{S}_m} - D^{il}_{\mathcal{S}_m}\right| \leq \frac{2}{m} + \frac{2}{m-n}$. This precise transductive setting allows the application of a bound as derived recently in [5], Lemma 2.

$$D_{\mathcal{S}_m} \leq E_{\mathcal{S}_m}[D_{\mathcal{S}_m}] + 2\sqrt{\left(\frac{1}{m} + \frac{1}{n-m}\right)\log(1/\delta)} \tag{6}$$

Now elaborating the term $E_{\mathcal{S}_m}[D_{\mathcal{S}_m}]$ gives the expression

$$E_{\mathcal{S}_m}\left[\sup_{q\in\mathcal{H}}\left(\mathcal{R}_{\mathcal{S}_{-m}}(q|\mathcal{G})-\mathcal{R}_{\mathcal{S}_m}(q|\mathcal{G})\right)\right]$$

$$= E_{\mathcal{S}_m}\left[\sup_{q\in\mathcal{H}}\frac{1}{m}\sum_{i\in\mathcal{S}_m} y_{\mathcal{S}_m(i)}q_{\mathcal{S}_m(i)} - \frac{1}{n-m}\sum_{i\notin\mathcal{S}_m} y_{\mathcal{S}_{-m}(i)}q_{\mathcal{S}_{-m}(i)}\right]$$

$$= E_{\mathcal{S}_m}\left[\sup_{q\in\mathcal{H}}\sum_{i=1}^{n}(y_iq_i)\left(I(i\in\mathcal{S}_m)\frac{1}{m} - I(i\in\mathcal{S}_{-m})\frac{1}{n-m}\right)\right], \quad (7)$$

where $I(i \in \mathcal{S}_m)$ equals one if $i \in \mathcal{S}_m$ and zero otherwise. As classical, introducing the Rademacher variables $\{\sigma_i\}_{i=1}^n$ which are independent from $y$ and $\mathcal{S}_m$, and satisfy $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$ does not change the expectation.

$$E_{\mathcal{S}_m,\sigma}\left[\sup_{q\in\mathcal{H}}\sum_{i=1}^{n}(y_iq_i)\,\sigma_i\left(I(i\in\mathcal{S}_m)\frac{1}{m} - I(i\in\mathcal{S}_{-m})\frac{1}{n-m}\right)\right]$$

$$\leq E_{\mathcal{S}_m,\sigma}\left[\sup_{q\in\mathcal{H}}\sum_{i=1}^{n}(y_iq_i)\,\sigma_i I(i\in\mathcal{S}_m)\frac{1}{m}\right] + E_{\mathcal{S}_m,\sigma}\left[\sup_{q\in\mathcal{H}}\sum_{i=1}^{n}(y_iq_i)\,\sigma_i I(i\in\mathcal{S}_{-m})\frac{1}{n-m}\right]$$

$$= 2\left(\frac{n}{n-m}\right)E_\sigma\left[\sup_{q\in\mathcal{H}}\sum_{i=1}^{n}q_i\sigma_i\right]. \quad (8)$$

This proofs the theorem.

□

## 3 Characterization of Plausible Sets $\mathcal{H}$

It becomes feasible to compute the measure $R(\mathcal{H}|\mathcal{G})$ empirically based on a Monte-Carlo sampling of the Rademacher variables. Since $\mathcal{H}$ is finite, one can find the $\sup_q$ for a given $\sigma \in \{-1,1\}^n$ by solving the problem

$$\hat{r}_\sigma = \max_{q\in\mathcal{H}}\frac{1}{n}q^T\sigma. \quad (9)$$

For many choices of $\mathcal{H}$, this combinatorial problem can be solved efficiently by an (approximative) algorithm. Averaging $\hat{r}_\sigma$ over the choice of $\sigma$ approximates the desired quantity. We will illustrate this for a relevant class $\mathcal{H}$.

### 3.1 $\mathcal{H}_1$ with Consistent Nearest Neighbors

**Definition 2 (1-NN Hypothesis Set)** *Assume $\mathcal{G}$ is connected, and that no two edges neighboring the same vertex have equal weight, or $x_{ij} \neq x_{ik}$ for all $i,j,k = 1,\ldots,n$. The 1-NN rule is defined as*

$$r_q^1(v_i) = q_{i^\star} \;\; \textit{with} \;\; i^\star = \arg\max_k x_{ik}. \quad (10)$$

*A labeling $q \in \{-1,1\}^n$ is consistent with this rule if $q_i r_q^1(v_i) = 1$ for all $i = 1,\ldots,n$ inducing the hypothesis space*

$$\mathcal{H}_1 = \left\{q\in\{-1,1\}^n \;\middle|\; q_i r_q^1(v_i) = q_i\, q_{i^\star} = 1,\;\; \forall i = 1,\ldots,n\right\}. \quad (11)$$

The main argument for characterizing $|\mathcal{H}_1|$ is to reduce a graph equipped with this rule to the maximal spanning tree $\mathcal{T}_{\text{MSP}}$. To make this clear, consider Kruskal's algorithm (see e.g. the excellent introduction [7]). Let $\mathcal{T} \subseteq \mathcal{G}$ be the current hypothesis, then Kruskal's algorithm effectively finds the maximal spanning tree as follows (1.) Find edge $e_{ij}$ with highest weight $x_{ij}$ in $E \setminus \mathcal{T}$; (2.) Do $\mathcal{T} = \mathcal{T} \cup e_{ij}$, if $\mathcal{T}$ is still a tree; (3.) Repeat 1 and 2 until $|\mathcal{T}| = (n-1)$. Now it is clear that the resulting maximal spanning tree $\mathcal{T}_{\text{MSP}}$ includes all edges $\{e_{i,i^\star}\}_{i=1}^n$ where $e_{i,i^\star}$ connects node $v_i$ to its closest neighbor $v_{(i)}$. Formally,

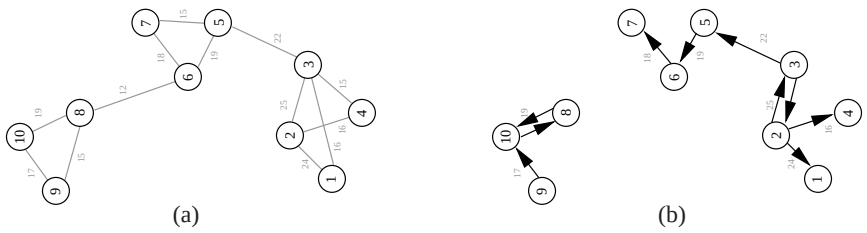

Figure 1: *(a) A weighted graph $\mathcal{G}$, and the compact representation $\mathcal{G}^1$ of the corresponding hypothesis space $\mathcal{H}_1$. By inspection of all $n$ directed edges in $\{e_{i,i^\star}\}_{i=1}^n$, one finds immediately that $\xi = 2$.*

**Lemma 1 (Maximal Spanning Tree and 1-NN)**

$$\{e_{i,i^\star}\}_{i=1}^n \subseteq \mathcal{T}_{MSP}. \tag{12}$$

Indeed, assume the edge $e_{i,i^\star}$ is rejected during some part of the algorithm, then there existed already some other edge $e_{ij}$ with higher weight $x_{ij}$ than $x_{i,i^\star}$, contradicting the assumption. Now a simple geometrical argument gives a characterization of the hypothesis space.

**Corollary 1 (Cardinality of $\mathcal{H}_1$)** *Let $\xi \in \mathbb{N}$ denote the number of disconnected parts in $\{e_{i,i^\star}\}_{i=1}^n$, or $\xi = |\mathcal{T}_{MSP} \setminus \{e_{i,i^\star}\}_i|$. Then the number of consistent (w.r.t. 1-NN rule) hypotheses $q$ equals*

$$\mathcal{H}_1 = 2^\xi. \tag{13}$$

This result follows from the observation that the omission of an edge in $\mathcal{T}_{\text{MSP}}$ results in one more disconnected set of vertices. Remark that this reasoning is conceptually different from the classical analysis of 1-NN rules (as found in e.g. [4], Chapter 5 and references). We are interested in the class of *consistent* labelings induced by the 1-NN rule, while in the classical account one studies the hypotheses induced by application of the 1-NN rule based on the observed labels.

An example is given in Figure 1. Given this representation, it is easy to select a hypothesis $q \in \mathcal{H}_1$ which makes the least amount of errors on the observed labels $\{y_i\}_{i \in \mathcal{S}_m}$ as possible, i.e.

$$\hat{q} = \arg\min_{q \in \mathcal{H}_1} \frac{1}{m} \sum_{i \in \mathcal{S}_m} y_i q_i, \tag{14}$$

by assigning to each disjoint subgraph $V' \subseteq V$ in $\mathcal{G}^1$ the label $\{-1, 1\}$ which occurs in $\{y_i\}_{(i \in \mathcal{S}_m, i \in V')}$. Note that the solution is not unique in the case the number of negative observed labels and positive labels in $V'$ equals (or is both zero). As a consequence of Lemma 3, one can immediately see that every graph $\mathcal{G}^1$ has at least one couple $(e_{i,i^\star}, e_{i^*,i})$.

### 3.2 $\mathcal{H}$ with Consistent Average Neighbors

For a qualitative discussion of the set $\mathcal{H}_g = \{q \in \{-1, 1\} : \ q^T L q \leq g\}$, its relation with maximal margin and spectral approximation, see [11, 9].

## 4 Conclusion

This note discusses some new results for transductive inference tasks defined over a single graph with respect to Rademacher complexities. For a specific hypothesis set, it is indicated how Kruskal's algorithm for finding the maximal (minimal) spanning tree gives the solution of the corresponding transductive inference task. It is clear that the choice of the set $\mathcal{H}$ is highly task dependent, and an important direction for future work concerns the data dependent choice (model selection) of a

relevant set $\mathcal{H}$, the exploration of other useful designs of $\mathcal{H}$ and the relation with graph algorithms. [1]

## References


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

[2] A. Blum and J. Langford. PAC-MDL bounds. In *Proceedings of the The Sixteenth Annual Conference on Learning Theory (COLT03)*, pages 344–357, 2003.

[3] O. Chapelle, B. Schölkopf, and A. Zien(Eds.), editors. *Semi-supervised Learning (In Press)*. MIT Press, Cambridge, MA, 2006.

[4] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

[5] R. El-Yaniv and D. Pechyony. Transductive rademacher complexity and its applications. In *Proceedings of the 20th Annual Conference on Computational Learning Theory (COLT)*, 2007.

[6] S. Hanneke. An analysis of graph cut size for transductive learning. In *In proceedings of the 23rd International Conference on Machine Learning (ICML).*, 2006.

[7] J. Kleinberg and E. Tardos. *Algorithmical Design*. Addison-Wesley, 2005.

[8] R. El-Yaniv P. Derbeko and R. Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Journal of Artificial Intelligence Research*, 22:117–142, 2004.

[9] K. Pelckmans, J. Shawe-Taylor, J.A.K. Suykens, and B. De Moor. Margin based transductive graph cuts using linear programming. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics, (AISTATS 2007), pp. 360-367*, San Juan, Puerto Rico, 2007.

[10] K. Pelckmans, J.A.K. Suykens, and B. De Moor. Transductive learning over graphs: Incremental assessment. In *International The Learning Workshop (SNOWBIRD), Technical Report ESAT-SISTA, K.U.Leuven (Leuven, Belgium), 2007-06*, San Juan, Puerto Rico, 2007.

[11] K. Pelckmans, J.A.K. Suykens, and B. De Moor. The kingdom-capacity of a graph: On the difficulty of learning a graph labelling. In *workshop on Mining and Learning with Graphs (MLG 2006)*, Berlin, Germany, 2006.

[12] V.N. Vapnik. *Statistical Learning Theory*. Wiley and Sons, 1998.



[1] **Acknowledgments**. Research supported by GOA AMBioRICS, CoE EF/05/006; (Flemish Government): (FWO): PhD/postdoc grants, projects, G.0407.02, G.0197.02, G.0141.03, G.0491.03, G.0120.03, G.0452.04, G.0499.04, G.0211.05, G.0226.06, G.0321.06, G.0553.06, G.0302.07. (ICCoS, ANMMM, MLDM); (IWT): PhD Grants,GBOU (McKnow), Eureka-Flite2 - Belgian Federal Science Policy Office: IUAP P5/22,PODO-II,- EU: FP5-Quprodis; ERNSI; - Contract Research/agreements: ISMC/IPCOS, Data4s, TML, Elia, LMS, Mastercard. JS is a professor and BDM is a full professor at K.U.Leuven Belgium.