Learning on graphs using Orthonormal Representation is Statistically Consistent

Rakesh S
Department of Electrical Engineering
Indian Institute of Science
Bangalore, 560012, INDIA
rakeshsmysore@gmail.com

Chiranjib Bhattacharyya
Department of CSA
Indian Institute of Science
Bangalore, 560012, INDIA
chiru@csa.iisc.ernet.in

Abstract

Existing research [4] suggests that embedding graphs on a unit sphere can be beneficial in learning labels on the vertices of a graph. However the choice of optimal embedding remains an open issue. Orthonormal representation of graphs, a class of embeddings over the unit sphere, was introduced by Lovász [2]. In this paper, we show that there exists orthonormal representations which are statistically consistent over a large class of graphs, including power law and random graphs. This result is achieved by extending the notion of consistency designed in the inductive setting to graph transduction. As part of the analysis, we explicitly derive relationships between the Rademacher complexity measure and structural properties of graphs, such as the chromatic number. We further show the fraction of vertices of a graph $G$, on $n$ nodes, that need to be labelled for the learning algorithm to be consistent, also known as labelled sample complexity, is $\Omega\left(\vartheta(G)^2/n\right)$ where $\vartheta(G)$ is the famous Lovász $\vartheta$ function of the graph. This, for the first time, relates labelled sample complexity to graph connectivity properties, such as the density of graphs. In the multiview setting, whenever individual views are expressed by a graph, it is a well known heuristic that a convex combination of Laplacians [7] tend to improve accuracy. The analysis presented here easily extends to Multiple graph transduction, and helps develop a sound statistical understanding of the heuristic, previously unavailable.

1 Introduction

In this paper we study the problem of graph transduction on a simple, undirected graph $G = (V, E)$, with vertex set $V = [n]$ and edge set $E \subseteq V \times V$. We consider individual vertices to be labelled with binary values, $\pm 1$. Without loss of generality we assume that the first $fn$ vertices are labelled, i.e., the set of labelled vertices is given by $S = [fn]$, where $f \in (0, 1)$. Let $\bar{S} = V \setminus S$ be the unlabelled vertex set, and let $y_S$ and $y_{\bar{S}}$ be the labels corresponding to subgraphs $S$ and $\bar{S}$ respectively.

Given $G$ and $y_S$, the goal of graph transduction is to learn predictions $\hat{y} \in \mathbb{R}^n$, such that $err_{\bar{S}}^{0:1}[\hat{y}] = \sum_{j \in \bar{S}} \mathbb{I}[y_j \neq \hat{y}_j]$, $\hat{y} = sgn(\hat{y})$ is small. To aid further discussion we introduce some notations.

Notation
Let $S^{n-1} = \{u \in \mathbb{R}^n ||u|| = 1\}$ denote a $(n - 1)$ dimensional sphere. Let $D_n$, $S_n$ and $S_n^+$ denote a set of $n \times n$ diagonal, square symmetric and square symmetric positive semidefinite matrices respectively. Let $\mathbb{R}^n_+$ be a non-negative orthant. Let $1_n \in \mathbb{R}^n$ denote a vector of all 1’s. Let $[n] := \{1, \ldots , n\}$. For any $M \in S_n$, let $\lambda_1(M) \geq \ldots \geq \lambda_n(M)$ denote the eigenvalues and $M_i$ denote the $i^{th}$ row of $M$, $\forall i \in [n]$. We denote the adjacency matrix of a graph $G$ by $A$. Let $d_i$ denote the degree of vertex $i \in [n]$, $d_i := A_{ii}1_n$. Let $D \in D_n$, where
Let $G$ denote the complement graph of $G$, with the adjacency matrix $A = I_n 1_n^\top - I - A$. For $K \in \mathbb{S}_n^+$, we denote $\omega(K, y) = \max_{\alpha \in \mathbb{R}^n_+} g(\alpha, K, y)\{ = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{j=1}^n \sum_{i=1}^n \alpha_i \alpha_j y_i y_j K_{ij}\}$. Let $\mathcal{Y} = \overline{\mathcal{Y}} = \{\pm 1\}$, $\hat{\mathcal{Y}} \subseteq \mathbb{R}$ be the label, prediction and soft-prediction spaces over $V$. Given a graph $G$ and labels $y \in \mathcal{Y}^n$ on $V$, let $\text{cat}(A, y) := \sum_{i,j \neq y_j} A_{ij}$. We use $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ to denote any loss function. In particular, for $a \in \mathcal{Y}$, $b \in \hat{\mathcal{Y}}$, let $\ell(a, b) = I[ab < 0]$, $\ell_{\text{hinge}}(a, b) = (1 - ab)_+$ and $\ell_{\text{ramp}}(a, b) = \min(1, (1 - ab)_+)$ denote the 0-1, hinge and ramp loss respectively. The notations $O, o, \Omega, \Theta$ will denote standard measures defined in asymptotic analysis [14].

Motivation

Regularization framework is a widely used tool for learning labels on the vertices of a graph [2, 3, 4]

$$\min_{\hat{y} \in \mathcal{Y}^n} \frac{1}{|S|} \sum_{i \in S} \ell(y_i, \hat{y}_i) + \lambda \hat{y}^\top K^{-1} \hat{y} \tag{1}$$

where $K$ is a kernel matrix and $\lambda > 0$ is an appropriately chosen regularization parameter. It was shown in [4] that the optimal $\hat{y}^*$ satisfies the following generalization bound

$$\mathbb{E}_S \left[\text{err}_S[\hat{y}^*]\right] \leq c_1 \inf_{\hat{y} \in \mathcal{Y}^n} \left\{ \text{err}_V[\hat{y}] + \lambda \hat{y}^\top K^{-1} \hat{y} \right\} + c_2 \left(\frac{\text{tr}_p(K)}{\lambda|S|}\right)^p$$

where $\text{err}_H[\hat{y}] := \frac{1}{|H|} \sum_{i \in H} \ell(y_i, \hat{y}_i)$, $H \subseteq V$.

In this paper, we formalize the notion of consistency of graph transduction algorithms and derive novel graph-dependent statistical estimates for the following formulation.

$$\Lambda_C(K, y_S) = \min_{\hat{y}_j \in \mathcal{Y}, j \in S} \min_{\alpha \in \mathbb{R}^n_+} \frac{1}{2} \alpha^\top K \alpha + C \sum_{i \in S} \ell(\hat{y}_i, y_i) + C \sum_{j \in S} \ell(\hat{y}_j, \bar{y}_j) \tag{2}$$

where $\hat{y}_j = \sum_{k \in S} K_{jk} y_k \alpha_i + \sum_{j \in S} K_{jk} y_j \alpha_j$, $\forall k \in V$. If all the labels are observed then [22] showed that the above formulation is equivalent to (1). We note that the normalization step considered by [4] is equivalent to finding an embedding of a graph on a sphere. Thus, we study orthonormal representations of graphs [2], which define a rich class of graph embeddings on an unit sphere. We show that the formulation (2), working with orthonormal representations of graphs is statistically consistent over a large class of graphs, including random and power law graphs. In the sequel, we apply Rademacher complexity to orthonormal representations of graphs and derive novel graph-dependent transductive error bound. We also extend our analysis to study multiple graph transduction. More specifically, we make the following contributions.

Contributions

The main contribution of this paper is that we show there exists orthonormal representations of graphs that are statistically consistent on a large class of graph families $G_c$. For a special orthonormal representation—LS labelling, we show consistency on Erdős Rényi random graphs. Given a graph $G \in G_c$, with a constant fraction of nodes labelled $f = O(1)$, we derive

---

1 $\lambda(a) = \max(a, 0)$.
2 We drop the argument $\hat{y}$, when implicit from the context.
3 Pure component is a connected subgraph, where all the nodes in the subgraph have the same label.
an error convergence rate of $\epsilon^\frac{0.1}{\delta^G} = O\left(\frac{\vartheta(G)}{n}\right)^{\frac{1}{2}}$, with high probability; where $\vartheta(G)$ is the Lovász $\vartheta$ function of the graph $G$. Existing work $[5]$ showed an expected convergence rate of $O\left(\sqrt{\frac{1}{n}}\right)$, however $q$ is dependent on the true labels of the unlabelled nodes. Hence their bound cannot be computed explicitly $[6]$. We also apply Rademacher complexity measure to the function class associated with orthonormal representations and derive a tight bound relating to $\vartheta(G)$, the chromatic number of the graph $G$. We show that the Laplacian inverse $[4]$ has $O(1)$ complexity on graphs with high connectivity, whereas LS labelling exhibits a complexity of $\Theta(n^\frac{1}{2})$. Experiments demonstrate superior performance of LS labelling on several real world datasets. We derive novel transductive error bound, relating to graph structural measures. Using our analysis, we show that observing labels of $\Omega\left(\frac{\vartheta(G)}{n}\right)^{\frac{1}{2}}$ fraction of the nodes is sufficient to achieve consistency. We also propose an efficient Multiple Kernel Learning (MKL) based algorithm, with generalization guarantees for multiple graph transduction. Experiments demonstrate improved performance in combining multiple graphs.

2 Preliminaries

Orthogonal Representation: $[2]$ introduced the idea of orthonormal representations for the problem of embedding a graph on a unit sphere. More formally, an orthonormal representation of a simple, undirected graph $G = (V, E)$ with $V = [n]$, is a matrix $U = [u_1, \ldots, u_n] \in \mathbb{R}^{d \times n}$ such that $u_i^T u_j = 0$ whenever $(i, j) \notin E$ and $u_i \in S^{d-1}$ for all $i \in [n]$.

Let $\text{Lab}(G)$ denote the set of all possible orthonormal representations of the graph $G$ given by $\text{Lab}(G) := \{ U \mid U$ is an Orthonormal Representation $\}$. $[1]$ recently introduced the notion of graph embedding to the Machine Learning community and showed connections to graph kernel matrices. Consider the set of graph kernels $\mathcal{K}(G) := \{ K \in S^+_n \mid K_{ii} = 1, \forall i \in [n] ; K_{ij} = 0, \forall (i, j) \notin E \}$. $[1]$ showed that for every valid kernel $K \in \mathcal{K}(G)$, there exists an orthonormal representation $U \in \text{Lab}(G)$; and it is easy to see the other way, $K = U^T U \in \mathcal{K}(G)$. Thus, the two sets, $\text{Lab}(G)$ and $\mathcal{K}(G)$ are equivalent. Orthonormal representation is associated with an interesting quantity, the Lovász number $[2]$, defined as: $\vartheta(G) = 2 \left( \min_{K \in \mathcal{K}(G)} \omega(K, 1_n) \right)$. $\vartheta$ function is a fundamental tool for combinatorial optimization and approximation algorithms for graphs.

Lovász Sandwich Theorem: $[2]$ Given an undirected graph $G = (V, E)$, $I (\hat{G}) \leq \vartheta (G) \leq \chi(G)$; where $I (\hat{G})$ is the independent number of the complement graph $\hat{G}$.

3 Statistical Consistency of Graph Transduction Algorithms

In this section, we formalize the notion of consistency of graph transduction algorithms. Given a graph $G_n = (V_n, E_n)$ of $n$ nodes, with labels of subgraph $S_n \subseteq V_n$ observable, let $\epsilon_{S_n} := \inf_{\hat{S} \in \hat{V}_n} \epsilon_{S_n}[\hat{y}]$ denote the minimal unlabelled node set error. Consistency is a measure of the quality of the learning algorithm $A$, comparing $\epsilon_{S_n}[\hat{y}]$ to $\epsilon_{S_n}^* \hat{y}$, where $\hat{y}$ are the predictions made by $A$. A related notion of loss consistency has been extensively studied in literature $[3][12]$, which only show that the difference $\epsilon_{S_n}[\hat{y}] - \epsilon_{S_n}^*[\hat{y}] \to 0$ as $n \to \infty$ $[6]$. This does not confirm the optimality of $A$, that is $\epsilon_{S_n}[\hat{y}] \to \epsilon_{S_n}^*[\hat{y}]$. Hence, a notion stronger than loss consistency is needed.

Let $G_n$ belong to a graph family $G$, $\forall n$. We define the $\ell$-regret, $\mathcal{R}(A) = \epsilon_{S_n}[\hat{y}] - \epsilon_{S_n}^*[\hat{y}]$ to be small. Since, the labelled nodes are drawn randomly, there is a small probability that one gets an unrepresentative subgraph $S_n$. However, for large $n$, we want $\ell$-regret to be close to zero with high probability. $[1]$ In other words, for every finite and fixed $n$, we want to have an estimate on the $\ell$-regret, which decreases as $n$ increases. We define the following notion of consistency of graph transduction algorithms to capture this requirement:

**Definition 1.** Let $G$ be a graph family and $f \in (0, 1)$ be fixed. Let $V = \{ (v_i, y_i, E_i) \}_{i=1}^{\infty}$ be an infinite sequence of labelled node set, where $y_i \in \mathcal{Y}$ and $E_i$ is the edge information of node $v_i$ with the previously observed nodes $v_1, \ldots, v_{i-1}$, $\forall i \geq 2$. Let $V_n$ be the first $n$ nodes in $V$, and let $G_n \sim G$.

\[\text{If } G \text{ is not deterministic (e.g., Erdös Réényi), then there is small probability that one gets an unrepresentative graph, in which case we want the } \ell \text{-regret to be close to zero with high probability over } G_n \sim G.\]
Let (2) be solved for $K$ to be greater than that of the Laplacian inverse on a large class of graphs. In particular, we study 4 Graph Complexity Measures

most of the proofs to the supplementary material

b) gather useful tools—

c) present results is consistent on a large class of graph families. To the best of our knowledge, we are not aware of any literature which provide an explicit empirical error convergence rate and prove consistency of the graph transduction algorithm considered. Before we prove our main result, we investigate

Theorem 4.1. Let $G = (V, E)$ be a simple, undirected graph with $V = [n]$, $U \in \text{Lab}(G)$ and $p \in [0, 1]$. Let $H_U = \{h | h = Ub, b \in \mathbb{R}^n, \|b\|_2 \leq tC\sqrt{n}\}, C > 0$, $t \in [0, 1]$ and let $K = U^\top U \in \mathcal{K}(G)$ be the graph-kernel corresponding to $U$. The Rademacher complexity of graph $G$ defined by $U$ is given by $R(H_U, p) = \frac{1}{n} \mathbb{E}_\sigma \left[ \sup_{h \in H_U} \sum_{i=1}^n \sigma_i \langle h, u_i \rangle \right]$.

The above definition is motivated from [9, 3]. This is an empirical complexity measure, suited for the transductive settings. We derive the following novel tight Rademacher bound

Theorem 4.1. Let $G = (V, E)$ be a simple, undirected graph with $V = [n]$, $U \in \text{Lab}(G)$ and $p \in [0, 1]$. Let $H_U = \{h | h = Ub, b \in \mathbb{R}^n, \|b\|_2 \leq tC\sqrt{n}\}, C > 0$, $t \in [0, 1]$ and let $K = U^\top U \in \mathcal{K}(G)$ be the graph-kernel corresponding to $U$. The Rademacher complexity of graph $G$ defined by $U$ is given by $R(H_U, p) = c_0 tC \sqrt{p\lambda_1(K)}$, where $1/2 \sqrt{2} \leq c_0 \leq \sqrt{2}$ is a constant.

The above result provides a lower bound for the Rademacher complexity for any unit sphere graph embedding. While upper-bounds maybe available [9, 3] but, to the best of our knowledge, this is the first attempt at establishing lower bounds. The use of orthonormal representations allow us to relate class complexity measure to graph-structural properties.

Corollary 4.2. For $C, t, p = O(1)$, $R(H_U, p) = O(\sqrt{\chi(G)})$. (Suppl.)

Such connections between learning theory complexity measures and graph properties was previously unavailable [9, 3]. Corollary 4.2 suggests that there exists graph regularizers with class complexity as large as $O(\sqrt{\chi(G)})$, which motivate us to find substantially better regularizers. In particular, we investigate LS labelling [16]; given a graph $G$, LS labelling $K_{LS} \in \mathcal{K}(G)$ is defined as

$$K_{LS} = A + I, \rho \geq |\lambda_n(A)|$$

LS labelling has high Rademacher complexity on a large class of graphs, in particular

Corollary 4.3. For a random graph $G(n, q), q \in [0, 1]$, where each edge is present independently w.p. $q$; for $C, t, q = O(1)$ the Rademacher complexity of the function class associated with LS labelling [5] is $\Theta(n^{1/4})$, with high probability. (Suppl.)
For the limiting case of complete graphs, we can show that Laplacian inverse \( L^{-1} \), the most widely used graph regularizer has \( O(1) \) complexity (Claim 2 Suppl.), thus indicating that it may be suboptimal for graphs with high connectivity. Experimental results illustrate our observation.

We derive a class complexity measure for unit sphere graph embeddings, which indicates the richness of the function class, and helps the learning algorithm to choose an effective embedding.

5 Generalization Error Bound

In the previous section, we applied Rademacher complexity to orthonormal representations. In this section we derive novel graph-dependent generalization error bounds, which will be used in Section 6. Following a similar proof technique as in \([3]\), we propose the following error bound—

**Theorem 5.1.** Given a graph \( G = (V, E) \), \( V = [n] \) with \( y \in \mathcal{Y}^n \) being the unknown binary labels over \( V \); let \( U \in \text{Lab}(G) \), and \( K \in \mathcal{K}(G) \) be the corresponding kernel. Let \( \mathcal{H}_U = \{ h | h = U\beta, \beta \in \mathbb{R}^n, \|\beta\|_\infty \leq C \} \), \( C > 0 \). Let \( \ell \) be any loss function, bounded in \([0, B]\) and \( L\)-Lipschitz in its second argument. For \( f \in (0, 1/2] \) let labels of subgraph \( S \subseteq V \) be observable, \( |S| = nf \). Let \( \bar{S} = V \setminus S \). For any \( \delta > 0 \) and \( h \in \mathcal{H}_U \), with probability \( \geq 1 - \delta \) over \( S \sim \Pi_f \)

\[
er_S[y] \leq er_S[y] + LC \sqrt{\frac{2\lambda_1(K)}{f(1 - f)}} + \frac{c_1 B}{1 - f} \sqrt{\frac{1}{nf} \log \frac{1}{\delta}}
\]

(4)

where \( y = U^\top h \) and \( c_1 > 0 \) is a constant. (Suppl.)

**Discussion**  Note that from \([2]\), \( \lambda_1(K) \leq \chi(G) \) and \( \chi(G) \) is in-turn bounded by the maximum degree of the graph \([21]\). Thus, if \( L, B, f = O(1) \), then for sparse, degree bounded graphs; for the choice of parameter \( C = \Theta(1/\sqrt{n}) \), the slack term and the complexity term goes to zero as \( n \) increases. Thus, making the bound useful. Examples include tree, cycle, path, star and \( d \)-regular (with \( d = O(1) \)). Such connections relating generalization error to graph properties was not available before. We exploit this novel connection to analyze graph transduction algorithms in Section 6. Also, in Section 7 we extend the above result to the problem of multiple graph transduction.

5.1 Max-margin Orthonormal Representation

To analyze \( e_{S}^{\vartheta^{-1}} \) relating to graph structural measure, the \( \vartheta \) function, we study the maximum margin induced by any orthonormal representation, in an oracle setting.

We study a fully ‘labelled graph’ \( G = (V, E, y) \), where \( y \in \mathcal{Y}^n \) are the binary labels on the vertices \( V \). Given any \( U \in \text{Lab}(G) \), the maximum margin classifier is computed by solving \( \omega(K, y) = g(\alpha^*, K, y) \) where \( K = U^\top U \in \mathcal{K}(G) \). It is interesting to note that knowing all the labels, the max-margin orthonormal representation can be computed by solving an SDP. More formally

**Definition 3.** Given a labelled graph \( G = (V, E, y) \), where \( V = [n] \) and \( y \in \mathcal{Y}^n \) are the binary labels on \( V \), let \( \mathcal{H} = \bigcup_{U \in \text{Lab}(G)} \mathcal{H}_U \), where \( \mathcal{H}_U = \{ h | h = U\beta, \beta \in \mathbb{R}^n \} \). Let \( K \in \mathcal{K}(G) \) be the kernel corresponding to \( U \in \text{Lab}(G) \). The max-margin orthonormal representation associated with the kernel function is given by \( K_{\text{mm}} = \text{argmin}_{K \in \mathcal{K}(G)} \omega(K, y) \).

By definition, \( K_{\text{mm}} \) induces the largest margin amongst other orthonormal representations, and hence is optimal. The optimal margin has interesting connections to the Lovász \( \vartheta \) function —

**Theorem 5.2.** Given a labelled graph \( G = (V, E, y) \), with \( V = [n] \) and \( y \in \mathcal{Y}^n \) being the binary labels on vertices. Let \( K_{\text{mm}} \) be as in Definition 3 then \( \omega(K_{\text{mm}}, y) = \vartheta(G)/2 \). (Suppl.)

Thus, knowing all the labels, computing \( K_{\text{mm}} \) is equivalent to solving the \( \vartheta \) function. However, in the transductive setting, \( K_{\text{mm}} \) cannot be computed. Alternatively, we explore LS labelling \([3]\), which gives a constant factor approximation to the optimal margin on a large class of graphs.

**Definition 4.** A class of labelled graphs \( \mathcal{G} = \{ G = (V, E, y) \} \) is said to be a Labelled SVM-\( \vartheta \) graph family, if there exist a constant \( \gamma > 1 \) such that \( \forall G \in \mathcal{G}, \omega(K_{\text{LS}}, y) \leq \gamma \omega(K_{\text{mm}}, y) \).

\(^{6}\)We can generalize our result for \( f \in (0, 1) \), but for the simplicity of the proof we assume \( f \in (0, 1/2] \). This is also true in practice, where the number of labelled examples is usually very small.
Let $\ell$ be the complement graph

We relate the maximum geometric margin induced by orthonormal representations to the $\vartheta$ function of the graph. This allows us to derive novel graph dependent learning theory estimates.

### 6 Consistency of Orthonormal Representation of Graphs

Aggregating results from Section 4 and 5, we show that Algorithm 1 working with orthonormal representations of graphs is consistent on a large class of graph families. For every finite and fixed $n$, we derive an estimate on $\varepsilon^{r^{0.1}}_{S_n}$.

**Theorem 6.1.** For the setting as in Definition 1, let $f \in (0, 1/2]$ be fixed. Let $\hat{y}$ be the predictions learnt by Algorithm 1 with inputs $U_n \in Lab(G_n)$, $y_{S_n}$, and $C^* = \left(\frac{\vartheta(G_n)(1-f)}{2n^2f\vartheta(G_n)}\right)^{\frac{1}{2}}$. Then $\exists U_n \in Lab(G_n)$, $\forall G_n$ such that with probability at least $1 - \frac{1}{n}$ over $S_n \sim \Pi_f$

$$
\varepsilon^{0.1}_{S_n}[\hat{y}] = O\left(\left(\frac{\vartheta(G_n)}{f^3(1-f)n}\right)^{\frac{1}{2}} + \frac{1}{1-f} \sqrt{\frac{\log n}{nf}}\right)
$$

**Proof.** Let $K_n \in K(G_n)$ be the max-margin kernel associated with $G_n$ (Definition 3), and let $U_n \in Lab(G)$ be the corresponding orthonormal representation. Since $r^{\text{ramp}}$ is an upper bound on $r^{0.1}$, we concentrate on bounding $\varepsilon^{r^{\text{ramp}}}_{S_n}[\hat{y}]$. Note that for any $C > 0$

$$
C|S_n| \cdot \varepsilon^{r^{\text{ramp}}}_{S_n}[\hat{y}] \leq C|S_n| \cdot \varepsilon^{\text{hinge}}_{S_n}[\hat{y}] \leq \Lambda_C(K_n, y_{S_n})
$$

$$
\leq \Lambda_C(K_n, y_n) = \frac{\vartheta(G_n)}{2}
$$

The last inequality follows from Theorem 5.2. Note that for ramp loss $L = B = 1$; using Theorem 5.1 for $\delta = \frac{1}{n}$, it follows that with probability at least $1 - \frac{1}{n}$ over the random draw of $S_n \sim \Pi_f$,

$$
\varepsilon^{r^{\text{ramp}}}_{S_n}[\hat{y}] \leq \frac{\vartheta(G_n)}{2Cnf} + C \sqrt{\frac{2\lambda_1(K_n)}{f(1-f)}} + c_1 \frac{\log n}{nf}
$$

where $c_1 = O(1)$. Using $\lambda_1(K_n) \leq \vartheta(G_n)$ and optimizing RHS for $C$, we get $C^* = \left(\frac{\vartheta^2(G_n)(1-f)}{2n^2f\vartheta(G_n)}\right)^{\frac{1}{2}}$. Plugging back $C^*$ and using $\vartheta(G_n)\vartheta(G_n) = n$, Theorem 6.1 proves the claim. \[\square\]

6 showed that $E_S[\varepsilon^{r^{0.1}}_{S_n}] = O(\sqrt{n})$. However, as noted in Section 4, the quantity $q$ is dependent on $S_n$, and hence their bounds cannot be computed explicitly [6].

We assume that the graph does not contain duplicate nodes with opposite labels, $\varepsilon^{r^{0.1}}_{S_n} \neq 0$. Thus, consistency follows from the fact that $\vartheta(G) \leq n$ and for large families of graphs it is $O(n^*)$ where $0 < c < 1$. This theorem points to the fact that if $f = O(1)$, then by Definition 1 Algorithm 1 is $r^{0.1}$-consistency over such class of graph families. Examples include

**Power-law graphs:** Graphs where the degree sequence follows a power law distribution. We show that $\vartheta(G) = O(\sqrt{n})$ for naturally occurring power law graphs (Claim 4 Suppl.). Thus, working with the complement graph $(\bar{G})$, makes Algorithm 1 consistent.
Random graphs: For $G(n, q)$ graphs, $q = O(1)$; with high probability $\vartheta(G(n, q)) = \Theta(\sqrt{n})$ [13]. Note that choosing $K_n$ for various graph families is difficult. Alternatively, for Labelled SVM-$\vartheta$ graph family (Definition 4), if Lovász $\vartheta$ function is sub-linear, then for the choice of LS labelling, Algorithm 1 is $\ell^0$-consistent. Examples include Mixture of random graphs (Section 5.1). Furthermore, we analyze the fraction of labelled nodes to be observed, such that Algorithm 1 is consistent.

**Corollary 6.2** (Labelled Sample Complexity). Given a graph family $G_c$, such that $\vartheta(G_n) = O(n^c)$, $\forall G_n \in G_c$ where $0 \leq c < 1$. For $C = C^*$ as in Theorem 6.1 $\frac{1}{2} \left( \frac{\vartheta(G_n)}{n} \right)^{1/3 - \varepsilon}$, $\varepsilon > 0$ fraction of labelled nodes is sufficient for Algorithm 1 to be $\ell^0$-consistent w.r.t. $G_c$.

The proof directly follows from Theorem 6.1. As a consequence of the above result, we can argue that for sparse graphs ($\vartheta(G)$ is large) one would need a larger fraction of nodes labelled, but for denser graphs ($\vartheta(G)$ is small) even a smaller fraction of nodes being labelled suffices. Such connections relating sample complexity and graph properties was not available before.

To end this section, we discuss on the possible extensions to the inductive setting (Claim 5, Suppl.)—we can show that the uniform convergence of $err_S$ to $err_S$ in the transductive setting (for $f = 1/2$) is a necessary and sufficient condition for the uniform convergence of $err_S$ to the generalization error. Thus, the results presented here can be extended to the supervised setting. Furthermore, combining Theorem 5.1 with the results of [9], we can also extend our results to the semi-supervised setting.

### 7 Multiple Graph Transduction

Many real world problems can be posed as learning on multiple graphs [19 ?]. Existing algorithms for single graph transduction [10, 15] cannot be trivially extended to the new setting. It is a well known heuristic that taking a convex combination of Laplacian improves classification performance for single graph transduction [10, 15] cannot be trivially extended to the new setting. It is a well known heuristic that taking a convex combination of Laplacian improves classification performance for single graph transduction [10, 15] cannot be trivially extended to the new setting.

**Problem 1.** Given $G = \{G^{(1)}, \ldots, G^{(m)}\}$ a set of simple, undirected graphs $G^{(k)} = (V^{(k)}, E^{(k)})$, defined on a common vertex set $V = [n]$. Without loss of generality we assume that the first $fn$ vertices are labelled, i.e., the set of labelled vertices is given by $S = [fn]$, where $f \in (0, 1)$. Let $\bar{S} = V \setminus S$ be the unlabelled node set. Let $y_S, y_{\bar{S}}$ be the labels of $S, \bar{S}$ respectively. Given $G$ and labels $y_S$, the goal is to accurately predict the labels of $y_{\bar{S}}$.

Let $K = \{K^{(1)}, \ldots, K^{(m)}\}$ be the set of kernels corresponding to graphs $G$, $K^{(k)} \in K(G^{(k)}), \forall k \in [m]$. We propose the following MKL style formulation for multiple graph transduction

$$\Psi_C(K, y_S) = \min_{\eta \in \mathbb{R}^m_+ \mid \eta_1 = 1} \min_{\bar{y}_S \in \mathcal{Y}_S} \max_{\alpha \in \mathbb{R}^m_+ \mid \alpha \| \omega \leq C} g \left( \alpha, \sum_{k=1}^m \eta_k K^{(k)}, y_S, \bar{y}_S \right)$$

(6)

Extending our analysis from Section 5, we propose the following error bound

**Theorem 7.1.** For the setting as in Problem 1 let $f \in (0, 1/2]$ and $K = \{K^{(1)}, \ldots, K^{(m)}\}, K^{(k)} \in K(G^{(k)}), \forall k \in [m]$. Let $\alpha^*, \eta^*, \bar{y}^*_S$ be the solution to $\Psi_C(K, y_S)$ (6). Let $\hat{y} = \sum_{i=1}^m \eta_i K^{(k)}\bar{Y} \alpha^*$, where $\bar{Y} \in D_n$, $\bar{Y}_{ii} = y_i$ if $i \in S$, otherwise $\bar{y}^*_i$. Then, for any $\delta > 0$, with probability $\geq 1 - \delta$ over the choice of $S \subseteq V$ such that $|S| = nf$

$$err_{S}^{\theta^0, 1}[\hat{y}] \leq \frac{\tilde{\Psi}(K, y)}{Cnf} + C \sqrt{\frac{2\vartheta(G^{(0)})}{f(1-f)}} + \frac{c_1}{1-f} \sqrt{\frac{1}{nf} \log \frac{1}{\delta}}$$

where $c_1 = O(1)$, $\tilde{\Psi}(K, y) = \min_{k \in [m]} \omega(K^{(k)}, y)$ and $G^{(0)}$ is the union of graphs $G^{(k)}$ (Suppl.).

The above result gives us the ability for the first time to analyze generalization performance of multiple graph transduction algorithms. The expression $\tilde{\Psi}(K, y)$ suggests that combining multiple graphs should improve performance over considering individual graphs separately. Similar to Section 6.

---

1As in Theorem 5.1 we can generalize our results for $f \in (0, 1)$.

2$G^{(j)} = (V, E^{(j)})$, where $(i, j) \in E^{(j)}$ if edge $(i, j)$ is present in atleast one of the graphs $G^{(k)} \in G$, $k \in [m]$. 

---

7
we can show that if one of the graph families $G^{(l)}, l \in [m]$ of $\mathbb{G}$ obey $\partial(G^{(l)}_n) = O(n^c), 0 \leq c < 1$; $G^{(l)}_n \in G^{(l)}$, then there exists orthonormal representations $\mathbb{K}$, such that the MKL style algorithm optimizing for $\mathbb{K}$ is $\ell^0$-consistent over $\mathbb{G}$ (Claim 6 Suppl.). We can also show that combining graphs improves labelled sample complexity (Claim 7 Suppl.). This is a first attempt in developing a statistical understanding for the problem of multiple graph transduction.

8 Experimental results

We conduct two sets of experiments.

**Superior performance of LS labelling:** We use two datasets—similarity matrices* from [11] and RBF kernels* as similarity matrices for the UCI datasets [8]. We built an unweighted graph by thresholding the similarity matrices about the mean. Let $L = D - A$. For the regularized formulation (1), with $10\%$ of labelled nodes observable, we test four types of kernel matrices—LS labelling (LS-lab), $(\lambda_1 I + L)^{-1}$ (Un-Lap), $(\lambda_2 I + D^{-1/2}LD^{-1/2})^{-1}$ (N-Lap) and K-Scaling (KS-Lap) [4]. We summarize the results. Each entry is accuracy in $\%$ w.r.t. 0-1 loss, and the results were averaged over 100 iterations. Since we are thresholding by mean, the graphs have high connectivity. Thus, from Corollary [4,5] the function class associated with LS labelling is rich and expressive, and hence it outperforms previously proposed regularizers.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LS-lab</th>
<th>Un-Lap</th>
<th>N-Lap</th>
<th>KS-Lap</th>
</tr>
</thead>
<tbody>
<tr>
<td>AuralSonar*</td>
<td>76.5</td>
<td>68.1</td>
<td>66.7</td>
<td>69.2</td>
</tr>
<tr>
<td>Yeast-SW-5-7*</td>
<td>60.4</td>
<td>54.1</td>
<td>52.9</td>
<td>53.3</td>
</tr>
<tr>
<td>Yeast-SW-5-12*</td>
<td>78.6</td>
<td>61.2</td>
<td>60.5</td>
<td>64.3</td>
</tr>
<tr>
<td>Yeast-SW-7-12*</td>
<td>75.6</td>
<td>64.0</td>
<td>59.5</td>
<td>63.1</td>
</tr>
<tr>
<td>Diabetes†</td>
<td>73.1</td>
<td>68.3</td>
<td>68.6</td>
<td>68.5</td>
</tr>
<tr>
<td>Fourclass†</td>
<td>73.3</td>
<td>69.3</td>
<td>71.2</td>
<td>71.8</td>
</tr>
</tbody>
</table>

**Table 1: Superior performance of LS labelling.**

Each entry is accuracy in $\%$.

<table>
<thead>
<tr>
<th>Graph 1×2</th>
<th>1×3</th>
<th>1×4</th>
<th>2×3</th>
<th>2×4</th>
<th>3×4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aud</td>
<td>62.8</td>
<td>64.8</td>
<td>68.3</td>
<td>59.3</td>
<td>50.8</td>
</tr>
<tr>
<td>Vis</td>
<td>68.9</td>
<td>65.6</td>
<td>68.9</td>
<td>69.1</td>
<td>70.3</td>
</tr>
<tr>
<td>Ttx</td>
<td>68.7</td>
<td>59.2</td>
<td>64.8</td>
<td>64.6</td>
<td>60.9</td>
</tr>
<tr>
<td>Unn</td>
<td>69.7</td>
<td>60.3</td>
<td>52.7</td>
<td>62.7</td>
<td>67.4</td>
</tr>
<tr>
<td>Maj</td>
<td>72.7</td>
<td>75.2</td>
<td>80.5</td>
<td>65.4</td>
<td>62.6</td>
</tr>
<tr>
<td>Int</td>
<td>80.6</td>
<td>83.6</td>
<td>86.0</td>
<td>90.9</td>
<td>75.3</td>
</tr>
<tr>
<td>MV</td>
<td>98.9</td>
<td>93.4</td>
<td>95.6</td>
<td>97.7</td>
<td>87.7</td>
</tr>
</tbody>
</table>

**Table 2: Multiple Graphs Transduction.**

Each entry is accuracy in $\%$.

9 Conclusion

For the problem of graph transduction, we show that there exists orthonormal representations that are consistent over a large class of graphs. We also note that the Laplacian inverse regularizer is suboptimal on graphs with high connectivity, and alternatively show that LS labelling is not only consistent, but also exhibits high Rademacher complexity on a large class of graphs. Using our analysis, we also develop a sound statistical understanding of the improved classification performance in combining multiple graphs.

---

9 Relevant resources at: mllab.csa.iisc.ernet.in/rakeshs/nips14

The $(i,j)^{th}$ entry of an RBF kernel is given by $\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. We set $\sigma$ to the mean distance.

11 Majority graph is a graph where an edge $(i,j)$ is present, if a majority of the graphs have the edge $(i,j)$. 

10 The $(i,j)^{th}$ entry of an RBF kernel is given by $\exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$. We set $\sigma$ to the mean distance.
References


