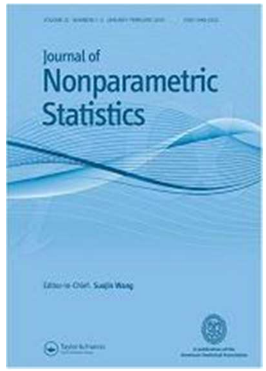

On Consistency of Graph-based Semi-supervised Learning

Journal:	*Journal of Nonparametric Statistics*
Manuscript ID	GNST-2017-06-10
Manuscript Type:	Original Paper
Date Submitted by the Author:	19-Jun-2017
Complete List of Authors:	Du, Chengan; George Mason University Volgenau School of Information Technology and Engineering Zhao, Yunpeng; George Mason University Volgenau School of Information Technology and Engineering, Statistics
Keywords:	Consistency, Semi-supervised learning, Graph Laplacian
Subject Area:	Semi-supervised learning, Nonparametric regression
2010 Mathematics Subject Classification:	62G08

To appear in the *Journal of Nonparametric Statistics*
Vol. 00, No. 00, Month 20XX, 1–15

On Consistency of Graph-based Semi-supervised Learning

Chengan Du[a] and Yunpeng Zhao[a]

[a]*Department of Statistics, George Mason University, VA, United States*

(*v4.0 released June 2015*)

Graph-based semi-supervised learning is one of the most popular methods in machine learning. Some of its theoretical properties such as bounds for the generalisation error and the convergence of the graph Laplacian regulariser have been studied in computer science and statistics literatures. However, a fundamental statistical property, the consistency of the estimator from this method has not been proved. In this article, we study the consistency problem under a non-parametric framework. We prove the consistency of graph-based learning in the case that the estimated scores are enforced to be equal to the observed responses for the labeled data. The sample sizes of both labeled and unlabeled data are allowed to grow in this result. When the estimated scores are not required to be equal to the observed responses, a tuning parameter is used to balance the loss function and the graph Laplacian regulariser. We give a counterexample demonstrating that the estimator for this case can be inconsistent. The theoretical findings are supported by numerical studies.

Keywords: Consistency ; Semi-supervised learning ; Graph Laplacian

1. Introduction

Semi-supervised learning is a class of machine learning methods that stand in the middle ground between supervised learning in which all training data are labeled, and unsupervised learning in which no training data are labeled. Specifically, in addition to the labeled training data $X_1, \ldots, X_n$, there exist unlabeled inputs $X_{n+1}, \ldots, X_{n+m}$. Under certain assumptions on the geometric structure of the input data, such as the cluster assumption or the low-dimensional manifold assumption (Chapelle, Schölkopf, and Zien 2006), the use of both labeled and unlabeled data can achieve better prediction accuracy than supervised learning that only uses labeled inputs $X_1, \ldots, X_n$.

Semi-supervised learning has become popular since the acquisition of unlabeled data is relatively inexpensive. A large number of methods were developed under the framework of semi-supervised learning. For example, Ratsaby and Venkatesh (1995) proposed that the combination of labeled and unlabeled data will improve the prediction accuracy under the assumption of mixture models. The self-training method (Rosenberg, Hebert, and Schneiderman 2005) and the co-training method (Jones 2005) were soon applied to semi-supervised learning when mixture models are not assumed. Zhang, Brady, and Smith (2001) described an approach to semi-supervised clustering based on hidden Markov random fields (HMRFs) that can combine multiple approaches in a unified probabilistic framework. Basu et al. (Basu, Banerjee, and Mooney. 2002) proposed a probabilistic framework for semi-supervised learning incorporating a K-means type hard partition

*Corresponding author. Email: latex.helpdesk@tandf.co.uk

clustering algorithm (HMRF-Kmeans). Vapnik (1998) proposed the transductive support vector machines (TSVMs) that used the idea of transductive learning by including unlabeled data in the computation of the margin. Transductive learning is a variant of semi-supervised learning which focuses on the inference of the correct labels for the given unlabeled data other than the inference of the general rule. Bie and Cristianini (2004) used a convex relaxation of the optimisation problem called semi-definite programming as a different approaches to the TSVMs.

In this article, we focus on a particular semi-supervised method – graph-based semi-supervised learning. In this method, the geometric structure of the input data are represented by a graph $\mathbf{G} = (\boldsymbol{V}, \mathbf{E})$, where nodes $\boldsymbol{V} = \{v_1, \ldots, v_{n+m}\}$ represent the inputs and edges $\mathbf{E}$ represent the similarities between them. The similarities are given in an $n+m$ by $n+m$ symmetric similarity matrix (or called *kernel* matrix), $\mathbf{W} = [w_{ij}]$, where $0 \leq w_{ij} \leq 1$. The larger w_{ij} implies that X_i and X_j are more similar. Further, let $Y_1, \ldots, Y_n$ be the responses of the labeled data.

Zhu, Ghahramani., and Lafferty. (2003) proposed the following graph-based learning method,

$$\min_{\boldsymbol{f}=(f_1,\ldots,f_{n+m})^T} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij}(f_i - f_j)^2 \tag{1}$$

$$\text{subject to } f_i = Y_i, i = 1, \ldots, n.$$

Its solution is called the estimated scores. The objective function (1) (named "hard criterion" thereafter), requires all the estimated score to be exactly the same as the responses for the labeled data. Delalleau et al. (Delalleau, Bengio, and Roux 2005) relaxed this requirement by proposing a soft version (named "soft criterion" thereafter). We follow an equivalent form given in Zhu and Goldberg (2009),

$$\min_{\boldsymbol{f}=(f_1,\ldots,f_{n+m})^T} \sum_{i=1}^{n} (Y_i - f_i)^2 + \frac{\lambda}{2} \sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij}(f_i - f_j)^2. \tag{2}$$

The soft criterion belongs to the "loss+penalty" paradigm: It searches for the minimiser $\hat{\boldsymbol{f}}$ which achieves a small training error, and in the meanwhile imposes the smoothness on $\hat{\boldsymbol{f}}$ by a penalty based on similarity matrix. It can be easily seen that when $\lambda = 0$ the soft criterion is equivalent to the hard criterion.

Remark 1 The tuning parameter λ being 0 in the soft criterion (2) is understood in the following sense: The squared loss has infinite weight and thereby $Y_i = f_i$ for all labeled data. But $\sum_{i=1}^{n+m} \sum_{j=1}^{n+m} w_{ij}(f_i - f_j)^2$ still plays a crucial role when it has no conflict with the hard constraints on the labeled data, that is, it provides links between f_i's on the labeled and unlabeled data. Therefore, the soft criterion (2) at $\lambda = 0$ becomes the hard criterion (1).

Zhou, Bousquet, Lal, Weston, and Schölkopf (2004); Belkin, Matveeva, and Niyogi (2004) have also proposed different variants of graph-based learning methods. We only focus on (1) and (2) in this article.

The theoretical properties of graph-based learning have been studied in computer science and statistics literatures. Bosquet, Chapelle, and Hein (2004) derived the limit of the Laplacian regulariser when the sample size of unlabeled data goes to infinity. Hein (2006) considered the convergence of Laplacian regulariser on Riemannian manifolds. Belkin, Niyogi, and Sindhwani (2006) reinterpreted the graph Laplacian as a measure of intrinsic distances between inputs on a manifold and reformulated the problem as a functional optimisation in a reproducing kernel Hilbert space. Nadler, Srebro, and Zhou (2009) pointed out that the hard criterion can yield completely noninformative solution when the size of unlabeled data goes to infinity and labeled data are finite, that is, the solution can give a perfect fit on the labeled data but remains as 0 on the unlabeled data. Lafferty and Wasserman (2008) obtained the asymptotic mean squared error of a different version of graph-based learning criterion. Belkin et al. (2004) gave a bound of the generalisation error for a slightly different version of (2). Alaoui, Cheng, Ramdas, Wainwright, and Jordan (2016) studied the theoretical properties of ℓ_p-based Laplacian regularisation – in particular the phase transition of p for a informative solution.

But to the best of our knowledge, no result is available in literature on a very fundamental question – the consistency of graph-based learning, which is the main focus of this article. Specifically, we want to answer the question that under what conditions $\hat{f}_i$ will converge to $\mathbb{E}[Y_i|X_i]$ on unlabeled data, where $\mathbb{E}[Y_i|X_i]$ is the true probability of a positive label given X_i if responses are binary, and $\mathbb{E}[Y_i|X_i]$ is the regression function on X_i if responses are continuous. We will always call $\mathbb{E}[Y_i|X_i]$ as regression function for simplicity.

Most of the literatures discussed above considered a "functional version" of (1) and (2). They used a functional optimisation problem with the optimiser $\hat{f}(x)$ being a function, as an approximation of the original problem with the optimiser $\hat{\boldsymbol{f}}$ being a vector. And they studied the behavior of the limit of graph Laplacian and the solution $\hat{f}(x)$. We do not adopt this framework but use a more direct approach. We focus on the original problem and study the relations of $\hat{f}_i$ and $\mathbb{E}[Y_i|X_i]$ under the general non-parametric setting. Our approach essentially belongs to the framework of transductive learning, which focuses on the prediction on the given unlabeled data $X_{n+1}, \ldots, X_{n+m}$, not the general mapping from inputs to responses. By establishing a link between the optimiser of (1) and the Nadaraya-Watson estimator (Nadaraya 1964; Watson 1964) for kernel regression, we will prove the consistency of the hard criterion. The theorem allows both m and n to grow. On the other hand, we show that the soft criterion is inconsistent for sufficiently large λ. To the best of our knowledge, this is the first result that explicitly distinguishes the hard criterion and the soft criterion of graph-based learning from a theoretical perspective and shows that they have very different asymptotic behaviors.

The rest of the article is organized as follows. In Section 2, we state the consistency result for the hard criterion and give the counterexample for the soft criterion. We prove the consistency result in Section 3. Numerical studies in Section 4 support our theoretical findings. Section 5 concludes with a summary and discussion of future research directions.

2. Main Results

We begin with basic notation and setup. Let $(X_1, Y_1), \ldots, (X_{n+m}, Y_{n+m})$ be independently and identically distributed pairs. Here each X_i is a d-dimensional vector and $\boldsymbol{Y} = (Y_1, \ldots, Y_{n+m})^T$ are binary responses labeled as 1 and 0 (the classification case) or continuous responses (the regression case). The last m responses are unobserved.

Zhu et al. (2003) used a fixed point algorithm to solve the hard criterion (1), which is

$$f_a = \frac{\sum_{i=1}^{n+m} w_{ak} f_i}{\sum_{i=1}^{n+m} w_{ai}}, \quad a = n+1, \ldots, n+m. \tag{3}$$

Note that (3) is not a closed-form solution but an updating formula for the iterative algorithm, since its right-hand side depends on unknown quantities.

In order to obtain a closed-form solution for (1), we begin by solving the soft version (2) and then let $\lambda = 0$. Recall that W is the similarity matrix. Let $\mathbf{D} = \text{diag}(d_1, \ldots, d_{n+m})$ where $d_i = \sum_{j=1}^{n+m} w_{ij}$, and $\mathbf{L} = \mathbf{D} - \mathbf{W}$ being the unnormalised graph Laplacian (see Newman (2010) for more details). Soft criterion (2) can be written in matrix form

$$\min_{\boldsymbol{f}} (\boldsymbol{f} - \boldsymbol{Y})^T \mathbf{V} (\boldsymbol{f} - \boldsymbol{Y}) + \lambda \boldsymbol{f}^T \mathbf{L} \boldsymbol{f}, \tag{4}$$

where $\mathbf{V}$ is an $n+m$ by $n+m$ matrix defined as

$$\mathbf{V} = \begin{pmatrix} \mathbf{I}_n & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}.$$

Then by taking the derivative of (4) with respect to $\boldsymbol{f}$ and setting equal to zero, we obtain the solution as follows,

$$\hat{\boldsymbol{f}} = (\mathbf{V} + \lambda \mathbf{L})^{-1} \mathbf{V} \begin{pmatrix} \boldsymbol{Y}_n \\ \mathbf{0} \end{pmatrix}.$$

where $\boldsymbol{Y}_n = (Y_1, \ldots, Y_n)^T$.

What we are interested in are the estimated scores on the unlabeled data, i.e. $\hat{\boldsymbol{f}}_{(n+1):(n+m)} = (\hat{f}_{n+1}, \ldots, \hat{f}_{n+m})^T$. In order to obtain an explicit form for $\hat{\boldsymbol{f}}_{(n+1):(n+m)}$, we use a formula for inverse of a block matrix (see standard textbooks on matrix algebra such as Intriligator and Griliches (1988) for more details): For any non-singular square matrix $\mathbf{A}$

$$\mathbf{A} = \begin{pmatrix} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \mathbf{A}_{21} & \mathbf{A}_{22} \end{pmatrix},$$

$$\mathbf{A}^{-1} = \begin{pmatrix} (\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1} & -(\mathbf{A}_{11} - \mathbf{A}_{12}\mathbf{A}_{22}^{-1}\mathbf{A}_{21})^{-1}\mathbf{A}_{12}\mathbf{A}_{22}^{-1} \\ -(\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1}\mathbf{A}_{21}\mathbf{A}_{11}^{-1} & (\mathbf{A}_{22} - \mathbf{A}_{21}\mathbf{A}_{11}^{-1}\mathbf{A}_{12})^{-1} \end{pmatrix}.$$

Write $\mathbf{D}$ and $\mathbf{W}$ as 2×2 block matrices,

$$\mathbf{D} = \begin{pmatrix} \mathbf{D}_{11} & \mathbf{D}_{12} \\ \mathbf{D}_{21} & \mathbf{D}_{22} \end{pmatrix}, \mathbf{W} = \begin{pmatrix} \mathbf{W}_{11} & \mathbf{W}_{12} \\ \mathbf{W}_{21} & \mathbf{W}_{22} \end{pmatrix}.$$

By the formula above,

$$\hat{\boldsymbol{f}}_{(n+1):(n+m)} = (\mathbf{D}_{22} - \mathbf{W}_{22} - \lambda \mathbf{W}_{21}(\mathbf{I}_n + \lambda \mathbf{D}_{11} - \lambda \mathbf{W}_{11})^{-1}\mathbf{W}_{12})^{-1}\mathbf{W}_{21}(\mathbf{I}_n + \lambda \mathbf{D}_{11} - \lambda \mathbf{W}_{11})^{-1}\boldsymbol{Y}_n. \quad (5)$$

By letting $\lambda = 0$, we obtain the solution for the hard criterion (1),

$$\hat{\boldsymbol{f}}_{(n+1):(n+m)} = (\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n. \quad (6)$$

Belkin et al. (2004) obtained a similar formula for a slightly different objective function.

Clearly, the form of (6) is closely related to the Nadaraya-Watson estimator (Nadaraya 1964; Watson 1964) for kernel regression, which is

$$\hat{q}_{n+a} = \frac{\sum_{i=1}^{n} w_{n+a,i} Y_i}{\sum_{k=1}^{n} w_{n+a,i}}, \quad a = 1, \ldots, m. \quad (7)$$

The Nadaraya-Watson estimator is well studied under the non-parametric framework. We can construct $\mathbf{W}$ by a kernel function, that is, let $w_{ij} = K((X_i - X_j)/h_n)$, where K is a nonnegative function on $\mathbb{R}^d$, and h_n is a positive constant controlling the bandwidth of the kernel. Let $q(X) = \mathbb{E}[Y|X]$ be the true regression function. The consistency of Nadaraya-Watson estimator was first proved by Watson (1964) and Nadaraya (1964). And many other researchers such as Devroye (1978) and Cai (2001) studied its asymptotic properties under different assumptions. Here we follow the result in Devroye and Wagner (1980). If $h_n \to 0$, $nh_n^d \to \infty$ as $n \to \infty$, and K satisfies:

(i) K is bounded by $k^* < \infty$;
(ii) The support of K is compact;
(iii) $K \geq \beta I_B$ for some $\beta > 0$ and some closed ball B centered at the origin and having positive radius δ,

then $\hat{q}_{n+a}$ converges to $q(X_{n+a})$ in probability for $a = 1, \ldots, m$.

By establishing a connection between the solution of the hard criterion and Nadaraya-Watson estimator, we prove the following main theorem:

Theorem 2.1 *Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n+m}, Y_{n+m})$ are independently and identically distributed with Y_i being bounded; h_n and K satisfy the above conditions. Further, we assume that the density function $\phi(\cdot)$ of X_1 has a compact support $\mathscr{X}$. And for every inner point x in $\mathscr{X}$,*

$$\phi(x) \geq s^* > 0. \quad (8)$$

Then, for $m = o(nh_n^d)$, $\hat{f}_{n+a}$ given in (5) converges to $q(X_{n+a})$ in probability, for $a = 1, \ldots, m$.

The proof will be given in Section 3.

Remark 2 Theorem 2.1 established the consistency of the hard criterion under the standard non-parametric framework with two additional assumptions. Firstly, both labeled data and unlabeled data are allowed to grow but the size of unlabeled data m grows slower than the size of labeled data n. We conjecture that when m grows faster than n, the graph-based semi-supervised learning may not be consistent based on the simulation studies in Section 4. Nadler et al. (2009) also suggested that the method may not work when m grows too fast. Secondly, we assume that density function of the difference of two independent inputs is strictly positive near the origin, which is a mild technical condition valid for commonly used density functions.

Theorem 2.1 provides some surprising insights about the hard criterion of graph-based learning. At a first glance, the hard criterion makes an impractical assumption that requires the responses to be noiseless, while the soft criterion seems to be a more natural choice. But according to our theoretical analysis, the hard criterion is consistent under the standard non-parametric framework where the responses on training data are of course allowed to be random and noisy.

We now consider the soft criterion with $\lambda \neq 0$.

Proposition 2.2 *Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots, (X_{n+m}, Y_{n+m})$ are independently and identically distributed with Y_i being bounded. Further, suppose that $\mathbf{W}$ represents a connected graph. Then for sufficiently large λ, the soft criterion (2) is inconsistent.*

Proof. Consider another extreme case of the soft criterion (2), $\lambda = \infty$. When $\mathbf{W}$ represents a connected graph, the objective function becomes

$$\min_{\boldsymbol{f}=(f_1,\ldots,f_n)^T} \sum_{i=1}^{n} (Y_i - f_i)^2 \tag{9}$$

$$\text{subject to } f_i = f_j, 1 \leq i, j \leq n+m.$$

It is easy to check that the solution of (9), denoted by $\hat{\boldsymbol{f}}(\infty)$, is given by

$$\hat{f}_{n+a}(\infty) = \frac{1}{n}\sum_{i=1}^{n} Y_i, \ a = 1, \ldots, m.$$

By the law of large numbers,

$$\lim_{n\to\infty} \hat{f}_{n+a}(\infty) = \mathbb{E}[q(X_1)] \text{ almost surely.}$$

Clearly, $\mathbb{E}[q(X_1)] \neq q(X_{n+a})$ since the right-hand side is a random variable. This implies that for sufficiently large λ, the soft criterion is inconsistent. ■

3. Proof of the Main Theorem

We give the proof of Theorem 2.1 in this section.

Recall that

$$\hat{\boldsymbol{f}}_{(n+1):(n+m)} = (\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n.$$

We first focus on $(\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}$. Clearly,

$$(\mathbf{D}_{22} - \mathbf{W}_{22})^{-1} = (\mathbf{I}_m - \mathbf{D}_{22}^{-1}\mathbf{W}_{22})^{-1}\mathbf{D}_{22}^{-1}.$$

For any positive integer l, define

$$\mathbf{S}_l = \mathbf{D}_{22}^{-1}\mathbf{W}_{22} + (\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^2 + (\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^3 + \cdots + (\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^l.$$

Our goal is to prove that the limit of S_l exists with probability approaching 1, and thus we can have

$$(\mathbf{I}_m - \mathbf{D}_{22}^{-1}\mathbf{W}_{22})^{-1} = \mathbf{I}_m + \lim_{l\to\infty} \mathbf{S}_l$$

with probability approaching 1 (Werner 2005).

By definition,

$$\mathbf{D}_{22} = \begin{pmatrix} d_{n+1,n+1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & d_{n+m,n+m} \end{pmatrix}, \mathbf{W}_{22} = \begin{pmatrix} w_{n+1,n+1} & \cdots & w_{n+1,n+m} \\ \vdots & \ddots & \vdots \\ w_{n+m,n+1} & \cdots & w_{n+m,n+m} \end{pmatrix},$$

where

$$d_{n+a,n+a} = \sum_{k=1}^{n+m} w_{n+a,k}, \quad w_{n+a,i} = K\left(\frac{X_i - X_{n+a}}{h_n}\right),$$

for $1 \le a \le m, 1 \le i \le n+m$. Thus we have

$$\mathbf{D}_{22}^{-1}\mathbf{W}_{22} = \begin{pmatrix} w_{n+1,n+1}/d_{n+1,n+1} & \cdots & w_{n+1,n+m}/d_{n+1,n+1} \\ \vdots & \ddots & \vdots \\ w_{n+1,n+m}/d_{n+m,n+m} & \cdots & w_{n+m,n+m}/d_{n+m,n+m} \end{pmatrix}.$$

Define $p(X_{n+a}) = \mathbb{P}(\|X_i - X_{n+a}\| \le \delta h_n \mid X_{n+a})$. Then since $h_n \to 0$, by the assumption in (8) and the definition of multiple integral, with probability 1.

$$\lim_{n\to\infty} \frac{p(X_{n+a})}{V_d(\delta h_n)} = \phi(X_{n+a}) \ge s^*,$$

where $V_d(\delta h_n)$ denotes the volume of a d-dimensional ball with radius δh_n. Then for sufficiently large n,

$$p(X_{n+a}) \ge \frac{1}{2} s^* V_d(\delta h_n) = s h_n^d,$$

where s is a constant only related to s^* and δ.

Since $nh_n^d \to \infty$, the above inequality implies $np(X_{n+a}) \to \infty$. On the other side, $p(X_{n+a}) \to 0$ since $h_n \to 0$.

Further,

$$\mathrm{Var}(I\{\|X_i - X_{n+a}\| \le \delta h_n\} \mid X_{n+a}) = p(X_{n+a})(1 - p(X_{n+a})).$$

By Chebyshev's Inequality, for any $0 < \epsilon < 1/2$, since $nh_n^d \to \infty$,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\}}{np(X_{n+a})} - 1\right| \ge \epsilon \middle| X_{n+a}\right) \\
=&\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\} - p(X_{n+a})\right| \ge \epsilon p(X_{n+a}) \middle| X_{n+a}\right) \\
\le&\frac{p(X_{n+a})(1-p(X_{n+a}))}{n\epsilon^2 p(X_{n+a})^2} \le \frac{1}{\epsilon^2 p(X_{n+a})n} \le \frac{1}{\epsilon^2 snh_n^d}. \qquad (10)
\end{aligned}$$

Therefore,

$$\mathbb{P}\left(\left|\frac{\sum_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\}}{np(X_{n+a})} - 1\right| \ge \epsilon\right) \le \frac{1}{\epsilon^2 snh_n^d} \to 0 \quad \text{as } n \to \infty. \qquad (11)$$

This further implies

$$\frac{\sum\limits_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\}}{np(X_{n+a})} \to 1 \quad \text{in probability.}$$

We now continue to study the property of $\mathbf{D}_{22}^{-1}\mathbf{W}_{22}$. Consider each element $(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})_{ab}$ of this matrix. For $1 \le a, b \le m$,

$$\begin{aligned}
(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})_{ab} &= \frac{w_{n+a,n+b}}{d_{n+a,n+a}} = K\left(\frac{X_{n+b} - X_{n+a}}{h_n}\right) \Big/ \sum_{i=1}^{n+m} K\left(\frac{X_i - X_{n+a}}{h_n}\right) \\
&\le \frac{k^*}{\beta \sum\limits_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\}},
\end{aligned}$$

by condition (i) and (iii). For simplicity of notation, let

$$\Phi_n(a) = \frac{\sum\limits_{i=1}^n I\{\|X_i - X_{n+a}\| \le \delta h_n\}}{np(X_{n+a})},$$

where Φ_n is a nonnegative function depending on n. By (10), we have

$$\mathbb{P}(0 \le \Phi_n(a) \le 1 - \epsilon) \le \mathbb{P}(|\Phi_n(a) - 1| \ge \epsilon) \le \frac{1}{\epsilon^2 snh_n^d},$$

which implies

$$\begin{aligned}
\mathbb{P}\left(\min_{1 \le a \le m} \Phi_n(a) \le 1 - \epsilon\right) &= \mathbb{P}\left(\bigcup_{a=1}^m \{\Phi_n(a) \le 1 - \epsilon\}\right) \\
&\le \sum_{a=1}^m \mathbb{P}(\Phi_n(a) \le 1 - \epsilon) \le \frac{m}{\epsilon^2 snh_n^d},
\end{aligned}$$

and

$$\mathbb{P}\left(\max_{1\le a\le m}\frac{k^*}{\beta\Phi_n(a)np(X_{n+a})}\le\frac{k^*}{\beta(1-\epsilon)np(X_{n+a})}\right)\ge 1-\frac{m}{\epsilon^2 snh_n^d}.$$

Since $\frac{m}{\epsilon^2 snh_n^d}\to 0$, we have

$$\mathbb{P}\left(\max_{1\le a,b\le m}(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})_{ab}\le\max_{1\le a\le m}\frac{k^*}{\beta\Phi_n(a)np(X_{n+a})}\le M\frac{1}{nh_n^d}\right)\to 1,\quad \text{as } n\to\infty, \quad (12)$$

where $M=\frac{2k^*}{s\beta}>\frac{k^*}{(1-\epsilon)s\beta}$. Note that M is a constant independent with n and m.

For the sake of simplicity, we say a matrix $\mathbf{A}$ has *tiny elements*, if

$$\|\mathbf{A}\|_{\max}\le M\frac{1}{nh_n^d},$$

with probability approaching 1, where $\|\mathbf{A}\|_{\max}=\max_{ij}\mathbf{A}_{ij}$. And $(\mathbf{A})_i$ denotes the i-th row of $\mathbf{A}$. Then $\mathbf{D}_{22}^{-1}\mathbf{W}_{22}$ has tiny elements by (12). Moreover,

$$\begin{aligned}\|(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^2\|_{\max}&=\|(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})\|_{\max}\\&\le(M\frac{1}{nh_n^d})^2m=\frac{M}{nh_n^d}(\frac{mM}{nh_n^d})\end{aligned}$$

holds with probability approaching 1. By induction,

$$\|(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^l\|_{\max}=\|(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^{l-1}\|_{\max}\le\frac{M}{nh_n^d}(\frac{mM}{nh_n^d})^{l-1},$$

with probability approaching 1. Therefore,

$$\begin{aligned}\|\mathbf{S}_l\|_{\max}=&\|\mathbf{D}_{22}^{-1}\mathbf{W}_{22}+\cdots+(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^l\|_{\max}\\\le&\|\mathbf{D}_{22}^{-1}\mathbf{W}_{22}\|_{\max}+\cdots+\|(\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^l\|_{\max}\\\le&\frac{M}{nh_n^d}\left(1+\cdots+(\frac{mM}{nh_n^d})^{l-1}\right)\ \text{with probability approaching 1.}\end{aligned}$$

$$\begin{aligned}\lim_{l\to\infty}\|\mathbf{S}_l\|_{\max}&\le\lim_{l\to\infty}\frac{M}{nh_n^d}\left(1+\cdots+(\frac{mM}{nh_n^d})^{l-1}\right)\\&\le\frac{M}{nh_n^d}/(1-\frac{mM}{nh_n^d})\le\frac{2M}{nh_n^d}\ \text{with probability approaching 1.}\end{aligned}$$

Thus $\mathbf{S}\triangleq\lim_{l\to\infty}\mathbf{S}_l$ exists with probability approaching 1 since $\lim_{l\to\infty}\|\mathbf{S}_l\|_{\max}<\infty$, and $\mathbf{S}$ also has tiny elements. Therefore,

$$(\mathbf{D}_{22}-\mathbf{W}_{22})^{-1}=(\mathbf{I}_m-\mathbf{D}_{22}^{-1}\mathbf{W}_{22})^{-1}\mathbf{D}_{22}^{-1}=(\mathbf{I}_m+\mathbf{S})\mathbf{D}_{22}^{-1},$$

with probability approaching 1.

We now go back to the solution of the hard criterion of graph-based semi-supervised learning,

$$\begin{aligned}\hat{\boldsymbol{f}}_{(n+1):(n+m)} &= (\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n \\ &= (\mathbf{I}_m + \mathbf{S})\mathbf{D}_{22}^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n = \mathbf{D}_{22}^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n + \mathbf{S}\mathbf{D}_{22}^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n, \end{aligned} \tag{13}$$

with probability approaching 1. For $1 \le a \le m$, $\hat{f}_{(n+a)}$ equals to the ath row of $(\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n$, i.e.,

$$\begin{aligned}\hat{f}_{(n+a)} &= \left((\mathbf{D}_{22} - \mathbf{W}_{22})^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n\right)_a \\ &= \sum_{i=1}^{n} \frac{w_{i,n+a}}{d_{n+a,n+a}} Y_i + (\mathbf{S})_a \mathbf{D}_{22}^{-1}\mathbf{W}_{21}\boldsymbol{Y}_n, \end{aligned} \tag{14}$$

with probability approaching 1, where $(\mathbf{S})_a$ denotes the ath row of $\mathbf{S}$.

By assumption, Y_i's are bounded. Without loss of generality, assume $\|Y_n\|_{\max} \le 1$. For $1 \le a \le m$, define

$$g_{(n+a)} = \sum_{i=1}^{n} Y_i \left(\frac{w_{i,n+a}}{\sum_{k=1}^{n} w_{k,n+a}} - \frac{w_{i,n+a}}{d_{n+a,n+a}} \right).$$

We have

$$\begin{aligned}|g_{(n+a)}| &\le \sum_{i=1}^{n} \|Y_n\|_{\max} \left(\frac{w_{i,n+a}}{\sum_{k=1}^{n} w_{k,n+a}} - \frac{w_{i,n+a}}{d_{n+a,n+a}} \right) \\ &= \frac{\sum_{i=1}^{n} w_{i,n+a}}{\sum_{k=1}^{n} w_{k,n+a}} - \frac{\sum_{i=1}^{n} w_{i,n+a}}{\sum_{k=1}^{n+m} w_{k,n+a}} \\ &= \frac{\sum_{k=n+1}^{n+m} w_{k,n+a}}{d_{n+a,n+a}} \\ &\le \frac{mk^*}{\beta\Phi_n(a) n p(X_{n+a})} \le \frac{mM}{nh_n^d} \to 0, \end{aligned}$$

with probability approaching 1 as $n \to \infty$. This implies

$$g_{(n+a)} \to 0 \text{ in probability},$$

since for any $\epsilon > 0$ we can find $m, n \in \mathbb{N}$ such that $\frac{mM}{nh_n^d} \le \epsilon$ and

$$\mathbb{P}(|g_{(n+a)}| \le \epsilon) \ge \mathbb{P}\left(|g_{(n+a)}| \le \frac{mM}{nh_n^d}\right) \to 1.$$

Finally, for each $1 \le a \le m$,

$$\hat{f}_{(n+a)} = \sum_{i=1}^{n} \frac{w_{i,n+a}}{d_{n+a,n+a}} Y_i + (\mathbf{S})_a \mathbf{D}_{22}^{-1} \mathbf{W}_{21} \boldsymbol{Y}_n$$
$$= \sum_{i=1}^{n} \frac{w_{i,n+a}}{\sum_{k=1}^{n} w_{k,n+a}} Y_i + (\mathbf{S})_a \mathbf{D}_{22}^{-1} \mathbf{W}_{21} \boldsymbol{Y}_n - g_{(n+a)},$$

Since $\mathbf{S}$ has tiny elements,

$$\|(\mathbf{S})_a \mathbf{D}_{22}^{-1} \mathbf{W}_{21} \boldsymbol{Y}_n\| \le \frac{mM}{nh_n^d} \to 0 \text{ with probability approaching 1,}$$

which implies $(\mathbf{S})_a \mathbf{D}_{22}^{-1} \mathbf{W}_{21} \boldsymbol{Y}_n \to 0$ in probability. The theorem then holds by the consistency of Nadaraya-Watson estimator.

4. Numerical Studies

In this section, we compare the performance of the hard criterion and the soft criterion with different tuning parameters under a linear and non-linear model.

The inputs $X_1, \ldots X_{n+m}$ are generated independently from a truncated multivariate normal distribution. Specifically, let $\tilde{X}_i$ follow a p-dimensional multivariate normal with the mean $\mu = (0.5, \ldots, 0.5)$, and the variance-covariance matrix

$$\begin{pmatrix} 0.1 & 0.05 & 0.05 & \ldots & 0.05 \\ 0.05 & 0.1 & 0.05 & \ldots & 0.05 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0.05 & 0.05 & 0.05 & \ldots & 0.1 \end{pmatrix}.$$

We set $p = 5$. For $i = 1, \ldots, n + m$ and $k = 1, \ldots, p$, let $X_{ik} = \tilde{X}_{ik}$ if $\tilde{X}_{ik} \in [0, 1]$ and $\tilde{X}_{ik} = 0$ otherwise, where X_{ik} and $\tilde{X}_{ik}$ are the k-th component of X_i and $\tilde{X}_i$, respectively.

Let $\mathbf{W}$ be the Gaussian radial basis function (RBF) kernel, that is,

$$w_{ij} = \exp\left(-\frac{\|X_i - X_j\|^2}{\sigma^2}\right), \text{ for } 1 \le i, j \le m + n,$$

where $\sigma = h_n = (\log n/n)^{1/5}$. Note that $\mathbf{W}$ has compact support since X_i's are truncated, and the choice of h_n satisfies the condition in Theorem 2.1.

We consider two models in simulation studies. In Model 1, the responses Y_i's follow a logistic regression with

$$\text{logit } q(X_i) = -1.35 + 2X_{i1} - X_{i2} + X_{i3} - X_{i4} + 2X_{i5},$$

for $i = 1, \ldots, m + n$. Model 2 uses a non-linear logit function,

$$\text{logit } q(X_i) = -1.35 + 2X_{i1} - X_{i2} + X_{i3} - X_{i4} + 2X_{i5} + X_{i1}X_{i3} + X_{i2}X_{i4},$$

for $i = 1, \ldots, m + n$.

We compare the performance of graph-based learning methods with four different tuning parameters, $\lambda = 0, 0.01, 0.1$ and 5. The performance is measured by the root mean squared error (RMSE) on the unlabeled data, that is,

$$\sqrt{\frac{1}{m}\sum_{a=1}^{m}(q(X_{n+a}) - \hat{q}_{n+a})^2}.$$

Each simulation is repeated 1000 times and the average RMSEs are reported.

Figure 1 shows the RMSEs under Model 1 when the sample size of unlabeled data m is fixed as 30 and the sample size of labeled data $n = 10$, 30, 50, 100, 200, 300, 500, 800, 1000 and 1500. As n increases, the RMSEs of all methods decrease as expected. More importantly, the RMSE increases as λ increases. In particular, the hard criterion always outperforms the soft criterion, which is consistent with our theoretical results.

Figure 2 shows the RMSEs under Model 1 when n is fixed as 100 and $m = 30$, 60, 100, 300, 500 and 1000. As before, the RMSE always increases as λ increases. Moreover, the RMSEs of all methods increase as m increases, which suggests that the hard criterion may not be consistent when m grows faster than n. For the non-linear logit function, Figure 3 and 4 show the same patterns as in Figure 1 and 2, respectively, which also support our theoretical results.

5. Summary

In this article, we proved the consistency of graph-based semi-supervised learning when the tuning parameter of the graph Laplacian is zero (the hard criterion) and showed that the method can be inconsistent when the tuning parameter is nonzero (the soft criterion). Moreover, the numerical studies also suggest that the hard criterion outperforms the soft criterion in terms of the RMSE. These results provide a better understanding about the statistical properties of graph-based semi-supervised learning. Of course, the accuracy of prediction can be measured by other indicators such as the area under the receiver operating characteristic curve (AUC). The hard criterion may not always be the best choice in term of these indicators. Further theoretical properties such as rank consistency will be explored in future research. Moreover, we would also like to investigate the behavior of these methods when the unlabeled data grow faster than the label data.

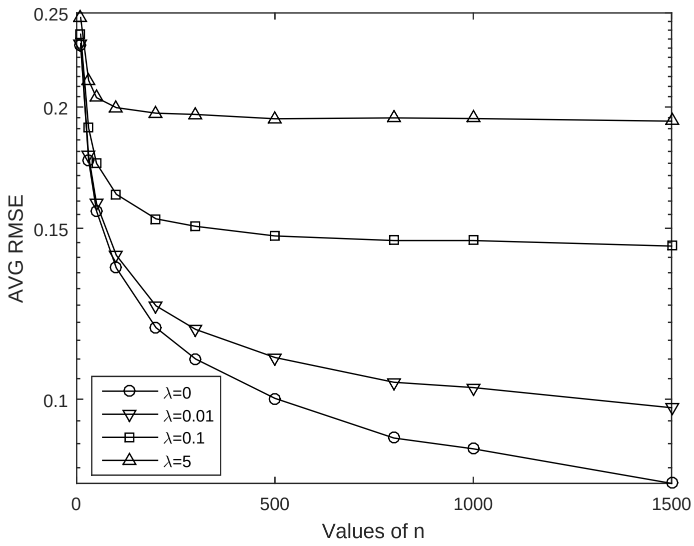

Figure 1.: Average RMSEs when $m = 30$ under Model 1

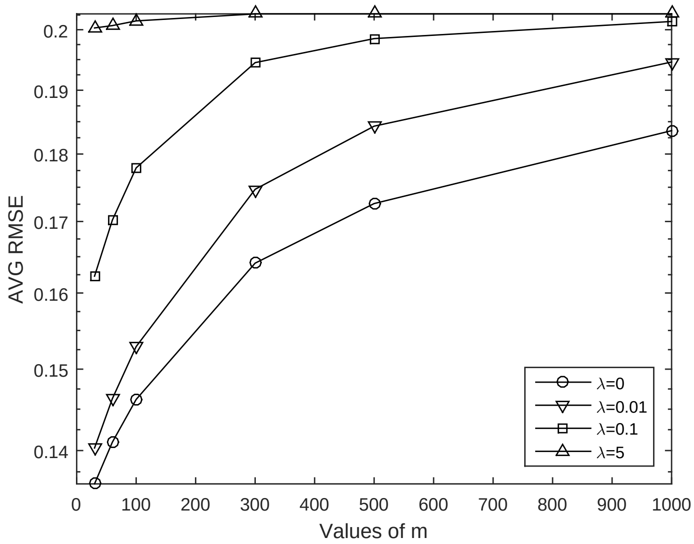

Figure 2.: Average RMSEs when $n = 100$ under Model 1

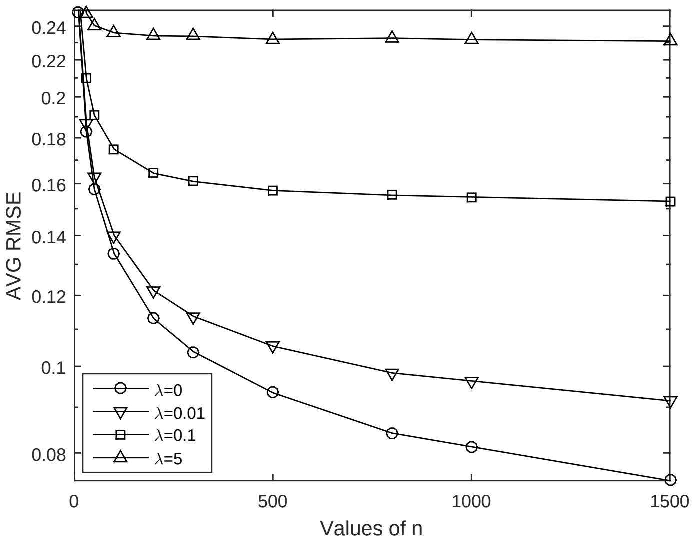

Figure 3.: Average RMSEs when $m = 30$ under Model 2

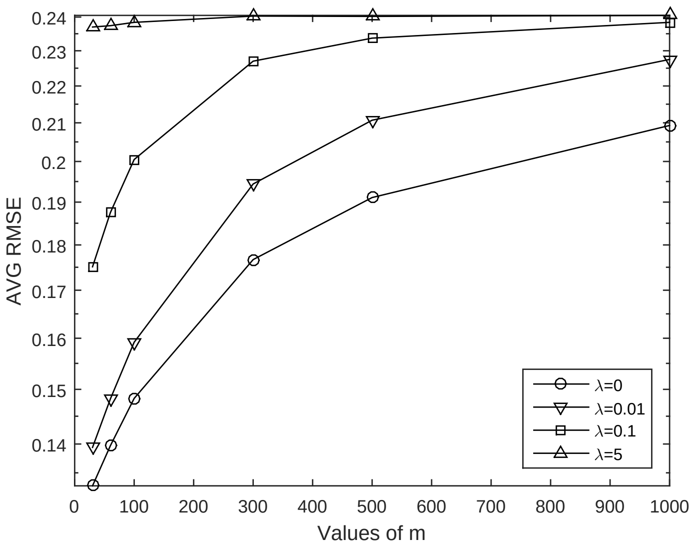

Figure 4.: Average RMSEs when $n = 100$ under Model 2

Funding

Yunpeng Zhao was supported by the U.S. National Science Foundation Grant DMS 1513004.

References

Alaoui, A.E., Cheng, X., Ramdas, A., Wainwright, M.J., and Jordan, M.I. (2016), 'Asymptotic behavior of ℓ_p-based Laplacian regularization in semi-supervised learning', in *29th Annual Conference on Learning Theory*, Vol. 49, Columbia University, New York, New York, USA, 23–26 Jun. Proceedings of Machine Learning Research, pp. 879–906.

Basu, S., Banerjee, A., and Mooney., R.J. (2002), 'Semi-supervised clustering by seeding', *In Proceedings of the International Conference on Machine Learning*, pp. 19–26.

Belkin, M., Matveeva, I., and Niyogi, P. (2004), 'Regularization and semi-supervised learning on large graphs', *In Proceedings of the Seventeenth Annual Conference on Computational Learning Theory*, pp. 624–638, Banff, Canada.

Belkin, M., Niyogi, P., and Sindhwani, S. (2006), 'Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples', *JMLR*, 7, 2399–2434.

Bie, T.D., and Cristianini, N. (2004), 'Convex methods for transduction', *In Advances in Neural Information Processing Systems*, 16, 73–80.

Bosquet, O., Chapelle, O., and Hein, M. (2004), 'Measure Based Regularization', *NIPS*, 16.

Cai, Z. (2001), 'Weighted Nadaraya Watson regression estimation', *Statistics & Probability Letters*, 51, 307–318.

Chapelle, O., Schölkopf, B., and Zien, A. (2006), *Semi-supervised Learning*, The MIT Press.

Delalleau, O., Bengio, Y., and Roux, N.L. (2005), 'Efficient Non-Parametric Function Induction in Semi-Supervised Learning', *In Artificial Intelligence and Statistics*.

Devroye, L.P. (1978), 'The uniform convergence of the Nadaraya-Watson regression function estimate', *The Canadian Journal of Statistics*, 6, 179–191.

Devroye, L.P., and Wagner, T.J. (1980), 'Distribution-free consistency results in nonparametric discrimination and regression function estimation', *Annals of Statistics*, 8, 231–239.

Hein, M. (2006), 'Uniform convergence of adaptive graph-based regularization', *COLT: Learning Theory*, pp. 50–64.

Intriligator, M., and Griliches, Z. (1988), *Handbook of Econometrics*, Vol. 1, North-Holland Publishing Company.

Jones, R. (2005), 'Learning to Extract Entitles from Labeled and Unlabeled Text', *PhD Thesis*.

Lafferty, J., and Wasserman, L. (2008), 'Statistical Analysis of Semi-Supervised Regression', *NIPS*, 20.

Nadaraya, E.A. (1964), 'On estimating regression', *Theor. Probability Appl*, 9, 141–142.

Nadler, B., Srebro, N., and Zhou, X. (2009), 'Semi-Supervised Learning with the Graph Laplacian: The Limit of Infinite Unlabelled Data', *NIPS*.

Newman, M.E.J. (2010), *Networks: An introduction*, Oxford University Press.

Ratsaby, J., and Venkatesh, S.S. (1995), 'Learning from a mixture of labeled and unlabeled examples with parametric side information', *In Proceedings of COLT '95 Proceedings of the eighth annual conference on Computational learning theory*, pp. 412–417.

Rosenberg, C., Hebert, M., and Schneiderman, H. (2005), 'Semi-Supervised Self-Training of Object Detection Models', *In Proceedings of Seventh IEEE Workshop on Applications of Computer Vision*.

Vapnik, V. (1998), *Statistical Learning Theory*, Wiley, New York.

Watson, G.S. (1964), 'Smooth regression analysis', *Sankhyā, Series A*, 26, 359–372.

Werner, D. (2005), *Functional Analysis (in German)*, Springer Verlag.

Zhang, Y., Brady, M., and Smith, S. (2001), 'Hidden Markov random field model and segmentation of brain MR images', *IEEE Transactions on Medical Imaging*, 20(1), 45–57.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., and Schölkopf, B. (2004), 'Learning with local and global consistency. In S. Thrun, L. Saul, and B. Schölkopf, editors', *MIT Press, Cambridge, MA*.

Zhu, X., and Goldberg, A.B. (2009), *Introduction to Semi-supervised Learning*, Morgan & Claypool Publishers.

Zhu, X., Ghahramani., Z., and Lafferty., J. (2003), 'Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions', *ICML*.