Statistical Inference on Graphs

Gérard BIAU * and Kevin BLEAKLEY

Institut de Mathématiques et de Modélisation de Montpellier
UMR CNRS 5149, Equipe de Probabilités et Statistique
Université Montpellier II, CC 051
Place Eugène Bataillon, 34095 Montpellier Cedex 5, France

biau,bleakley@math.univ-montp2.fr

Abstract

The problem of graph inference, or graph reconstruction, is to predict the presence or absence of edges between a set of given points known to form the vertices of a graph. Motivated by various applications including communication networks and systems biology, we propose a general model for studying the problem of graph inference in a supervised learning framework. In our setting, both the graph vertices and edges are assumed to be random, with a probability distribution that possibly depends on the size of the graph. We show that the problem can be transformed into one where we can use statistical learning methods based on empirical minimization of natural estimates of the reconstruction risk. Convex risk minimization methods are also studied to provide a theoretical framework for reconstruction algorithms based on boosting and support vector machines. Our approach is illustrated on simulated graphs.

Index Terms — Statistical learning, Classification, Vapnik-Chervonenkis dimension, Graph inference, Graph reconstruction, U-statistics, weakly convergent measures.

AMS 2000 Classification: 62G08, 68Q32, 05C80.

1 Introduction

Graph theory deals with problems that have a graph (or network) structure. In this context an (undirected) graph consists of:

*Corresponding author.

- vertices/nodes – which are a collection of points;
- and edges/arcs – which are non-oriented lines running between the vertices.

Graph and network structures such as the Internet or road networks appear all around us. Similarly, various biological networks such as gene and molecular networks can be modelled as graphs. Probabilistic modelling of graphs emerged in the second half of the 20th century in the works of P. Erdös and A. Rényi to study all kinds of structural properties of graphs, such as the emergence of cycles, as more and more edges are added randomly. We refer the reader to Bollobás [4] for a general overview. The present paper deals with the problem of *graph inference*, or *graph reconstruction*. The goal is to infer the presence or absence of edges between a set of *given* points known to form the vertices of a graph, given that we have some prior information about the vertices. This problem has recently received considerable attention in computational and systems biology, where the goal is to study the structural properties of various biological networks, such as protein interaction networks [13], gene regulatory networks [11] and metabolic networks [14]. To do this, a high quality (low noise) reconstruction of these networks is needed to begin with.

Various practical approaches to reconstruct networks have been proposed and tested so far, ranging from Bayesian or Petri networks to the prediction of edges between "similar" vertices. However, most approaches base their prediction on prior knowledge about which edges should be present for a given set of vertices. This prior knowledge might be for example based on a model of conditional dependence in the case of Bayesian networks, or the assumption that edges should connect similar vertices.

Recently, Vert and Yamanishi [24] pointed out that the actual situation we are confronted with, however, can often be expressed in a supervised learning framework. In it, we already have data about the vertices, and as well as that we already know part of the network, due to previous experimental results. The challenge then is to infer new edges involving particular vertices outside of the known subgraph. To solve the problem, Vert and Yamanishi [24] propose an algorithm which learns a mapping of the vertices into a Euclidean space, and then put an edge between the (mapped) vertices which are closer to each other than a threshold. The method can be formulated as a constrained optimization problem in a reproducing kernel Hilbert space.

Starting from the same point of view as Vert and Yamanishi [24], we wish to establish in this contribution a simple but rigorous and systematic statistical framework for studying the supervised graph inference problem. In our model, graph vertices are represented by nonatomic random variables, and the undirected edges (or lack of) are described by random binary labels: $+1$ if two vertices are connected, -1 otherwise. We discuss this general model and show that the reconstruction problem may be neatly transformed into one for which standard empirical risk minimization methods can be applied, with two important particularities.

- First, given a graph with n nodes, we allow the distribution of the nodes and $n(n-1)/2$ edges to depend on n, and we assume that this distribution stabilizes as n grows (in a sense to be made precise later). This model allows us to deal with networks that are not "scale free", such as the protein network controlling gene expression, whose operation is reliant on the integrated activity of many or all of its component nodes. Such networks are called "accelerating networks" (see Mattick and Gagen [17] for more information), as opposed to "scale free" networks such as the Internet, which is much more locally integrated than globally.

- Second, we point out that the natural estimates of the reconstruction risk involve data functionals which are close to, but not quite exactly, U-statistics. The introduction of U-statistics and related processes in statistical learning theory is recent, the systematic treatment of which goes back to Clémençon, Lugosi and Vayatis [6].

The paper is organized as follows. In Section 2, we set up notation and formally describe the problem of statistical inference on graphs. In Section 3, we provide the basic consistency results for empirical risk minimization. Section 4 is devoted to the study of reconstruction rules based on empirical minimization of convex cost functionals. The methods are inspired by boosting and support vector machine-type algorithms. Our approach is illustrated in Section 5 where we apply several of these reconstruction algorithms to simulated graphs.

2 The problem of statistical inference on graphs

Let n be an integer greater than or equal to 2. Throughout the paper, we denote by $\mathcal{X}$ a separable Banach space endowed with its Borel σ-field. We suppose the n vertices of a graph in $\mathcal{X}$ are represented by n *nonatomic*, *independent* and *identically distributed* $\mathcal{X}$-valued random variables $X_1^n, \ldots, X_n^n$.

The nonatomic requirement on the X_i^n's is to ensure that two vertices will almost surely never overlap. Then, given two vertices X_i^n and X_j^n, with $i \neq j$, the connection between these two vertices is modelled by a random binary label Y_{ij}^n, taking the value $+1$ if the nodes are connected, and -1 otherwise. For now we require for the sake of coherence that $Y_{ij}^n = Y_{ji}^n$ so as to study properties (between pairs of vertices) which do not represent a directional idea. This requirement means that the graph is undirected.

Thus, we consider a graph as a sequence of random variables (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i \neq j \leq n$, taking values in $\mathcal{X} \times \mathcal{X} \times \{-1, 1\}$. We note again that in this formalization, both the vertices and the edges between them are assumed to be random. As well as this, we assume from now on that the random variables (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i \neq j \leq n$, have the same distribution as a generic variable $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$, and also that the pairs (X_i^n, X_j^n, Y_{ij}^n) and (X_k^n, X_l^n, Y_{kl}^n) are independent for $\{i, j\} \cap \{k, l\} = \emptyset$. The superscript n in (X_i^n, X_j^n, Y_{ij}^n) and $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ may be a little confusing. It means that the distribution of the nodes and edges is allowed to depend on the network size n — in other words, we are faced with an *array* of $n(n-1)/2$ random variables (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$. As discussed in the introduction, this requirement is motivated by the observation that many real networks exhibit accelerating properties.

The probability distribution $P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}$ of $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ represents the "random state" of the graph with n vertices. Going one step further into modelling, we will assume throughout the paper that the following fundamental assumption holds:

Assumption 2.1. There exists an $\mathcal{X} \times \mathcal{X} \times \{-1, 1\}$-valued random variable $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$ such that $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ converges in distribution to $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$ as $n \to \infty$.

Let us denote by $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$ the probability distribution of the limit variable $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$. Assumption 2.1 is equivalent to the weak convergence of $P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}$ to $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$ (cf. Dudley [9] or Billingsley [1]), and, roughly speaking, tells us that the random state $P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}$ of the graph asymptotically stabilizes toward a limit state described by $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$. We believe that a large variety of dynamic behaviors can be captured by this model, including stability ($P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)} = P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$ for all $n \geq 2$) and acceleration.

In the graph inference problem, we observe a graph (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$, satisfying the requirements discussed above and sampled from

$P_{(\mathbb{X}_1^n,\mathbb{X}_2^n,\mathbb{Y}^n)}$. The aim is to design a (measurable) *reconstruction rule* $g : \mathcal{X}\times\mathcal{X}\to\{-1,1\}$ which predicts $\mathbb{Y}^n$ from $(\mathbb{X}_1^n,\mathbb{X}_2^n)$. The performance of such a rule is measured by the *reconstruction risk*:

$$\mathbf{R}^n(g)=\mathbb{P}\{g(\mathbb{X}_1^n,\mathbb{X}_2^n)\neq\mathbb{Y}^n\},$$

that is, the probability that g incorrectly predicts the presence (or absence) of an edge. Ideally, $\mathbf{R}^n(g)$ should be close to the Bayes risk, that is, to the minimal probability of error

$$\mathbf{R}_*^n=\inf_g\mathbf{R}^n(g)=\frac{1}{2}-\frac{1}{2}\mathbb{E}\big\{|2\eta^n(\mathbb{X}_1^n,\mathbb{X}_2^n)-1|\big\},\tag{2.1}$$

where the infimum is taken over all measurable reconstruction rules $g : \mathcal{X}\times\mathcal{X}\to\{-1,1\}$, and $\eta^n(x_1,x_2)=\mathbb{P}\big\{\mathbb{Y}^n=1|(\mathbb{X}_1^n,\mathbb{X}_2^n)=(x_1,x_2)\big\}$ denotes the posterior probability. It is well known (see e.g. Devroye, Györfi and Lugosi [8]) that the infimum in (2.1) is achieved by the Bayes classifier $g_*^n(x_1,x_2)=2\mathbf{1}_{\{\eta^n(x_1,x_2)>1/2\}}-1$. (Note that ties are broken in favor of -1, but this has no consequence on the value of the Bayes error).

As explained earlier, so as to consider undirected properties between pairs of vertices, we will limit ourselves to *symmetric* rules. Given such a rule g, we have to use the training graph to estimate the risk $\mathbf{R}^n(g)$. One natural estimate is the statistic

$$R_n(g)=\frac{2}{n(n-1)}\sum_{1\leq i<j\leq n}\mathbf{1}_{\{g(X_i^n,X_j^n)\neq Y_{ij}^n\}}.\tag{2.2}$$

This somewhat unusual definition of the empirical risk deserves several comments. We note first that R_n depends on n both through its empirical feature (it is a sum of $n(n-1)/2$ random variables) and through the sampling distribution $P_{(\mathbb{X}_1^n,\mathbb{X}_2^n,\mathbb{Y}^n)}$. Therefore, the notation R_n^n would have probably been more appropriate. However, for simplicity, and since no confusion is possible, we will continue to write R_n. On the other hand, using the fact that g is symmetric and $Y_{ij}^n=Y_{ji}^n$, we have

$$R_n(g)=\frac{1}{n(n-1)}\sum_{1\leq i\neq j\leq n}\mathbf{1}_{\{g(X_i^n,X_j^n)\neq Y_{ij}^n\}}.$$

It is worth mentioning that $R_n(g)$ is *not* a U-statistic (de la Peña and Giné [18]) *stricto sensu*, unless we assume the deterministic case $Y_{ij}^n=\eta^n(X_i^n,X_j^n)$. Consequently, the approach of Clémençon, Lugosi and Vayatis [6], who use U-statistics in a learning ranking problem, does not carry over.

3 Empirical risk minimization

3.1 Error bounds

A simple and natural approach to the graph inference problem is to consider choosing a reconstruction rule by minimizing the empirical risk $R_n(g)$ defined in (2.2) over a class $\mathcal{G}$ of candidate symmetric rules $g : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$. We denote by g_n a reconstruction rule which minimizes the empirical error over $\mathcal{G}$, given the data set (X_i^n, X_j^n, Y_{ij}^n), $1 \le i < j \le n$:

$$R_n(g_n) \le R_n(g) \quad \text{for all } g \in \mathcal{G}.$$

With a slight abuse of notation, we still denote by $\mathbf{R}^n(g_n)$ the conditional probability of error of the selected rule, that is

$$\mathbf{R}^n(g_n) = \mathbb{P}\{g_n(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n | (X_i^n, X_j^n, Y_{ij}^n), 1 \le i < j \le n\}.$$

The risk $\mathbf{R}^n(g_n)$ is easily seen to satisfy the elementary inequality (see Devroye, Györfi and Lugosi [8])

$$\mathbb{E}\mathbf{R}^n(g_n) - \inf_{g \in \mathcal{G}} \mathbf{R}^n(g) \le 2\mathbb{E}\left\{ \sup_{g \in \mathcal{G}} \left| R_n(g) - \mathbf{R}^n(g) \right| \right\}. \tag{3.1}$$

This means that if we can guarantee that the average uniform deviation

$$\mathbb{E}\left\{ \sup_{g \in \mathcal{G}} \left| R_n(g) - \mathbf{R}^n(g) \right| \right\}$$

of estimated probabilities from their true values is small, then we in turn ensure that the error probability of g_n is close to the best error probability over all $g \in \mathcal{G}$. The following lemma shows that, under our assumptions, we can reduce the network problem to the classical problem of bounding ordinary empirical processes. The proof is a slight adaptation of Lemma 1 in Clémençon, Lugosi and Vayatis [6], but the main arguments go back to Hoeffding [12], page 25.

Lemma 3.1 *Let Φ be a family of measurable functions*

$$\varphi : \mathcal{X} \times \mathcal{X} \times \{-1, 1\} \to \mathbb{R}$$

symmetric in the first two arguments. Then

$$\begin{aligned}
&\mathbb{E}\left\{ \sup_{\varphi \in \Phi} \frac{2}{n(n-1)} \left| \sum_{1 \le i < j \le n} \varphi(X_i^n, X_j^n, Y_{ij}^n) \right| \right\} \\
&\quad \le \mathbb{E}\left\{ \sup_{\varphi \in \Phi} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \varphi(X_i^n, X_{i+\lfloor n/2 \rfloor}^n, Y_{i\,i+\lfloor n/2 \rfloor}^n) \right| \right\},
\end{aligned}$$

where the notation $\lfloor . \rfloor$ stands for the integer part.

Under our basic assumptions, for each fixed function φ, we note that the term $\sum_{i=1}^{\lfloor n/2 \rfloor} \varphi(X_i^n, X_{i+\lfloor n/2 \rfloor}^n, Y_{i\,i+\lfloor n/2 \rfloor}^n)$ is just a sum of $\lfloor n/2 \rfloor$ *independent* and *identically distributed* random variables, and therefore all of the standard methods of dealing with empirical processes can be used directly. Hence the interest in this lemma, which we can now apply directly to our problem at hand. For example, taking

$$\varphi(X_i^n, X_j^n, Y_{ij}^n) = \mathbf{1}_{\{g(X_i^n, X_j^n) \neq Y_{ij}^n\}} - \mathbb{E}\mathbf{1}_{\{g(X_i^n, X_j^n) \neq Y_{ij}^n\}},$$

and assuming that $\mathcal{G}$ is a Vapnik-Chervonenkis class (Vapnik and Chervonenkis [23]), we immediately obtain the following corollary:

Corollary 3.1 *Let $\mathcal{G}$ be a class of symmetric reconstruction rules of finite Vapnik-Chervonenkis dimension $V_{\mathcal{G}}$. Then*

$$\mathbb{E}\left\{ \sup_{g \in \mathcal{G}} \left| R_n(g) - \mathbf{R}^n(g) \right| \right\} \leq C \sqrt{\frac{V_{\mathcal{G}}}{n}},$$

where C is a positive universal constant.

Lemma 3.1 and Corollary 3.1 provide us with simple and convenient tools that allow us to reduce the study of the empirical reconstruction risk R_n (2.2) to a more classical problem.

3.2 Asymptotic risk behavior

Given a graph (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$, sampled from the probability distribution $P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}$, the results of Paragraph 3.1 are summarized by the performance inequality:

$$\mathbb{E}\mathbf{R}^n(g_n) - \inf_{g \in \mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\} \leq 2C \sqrt{\frac{V_{\mathcal{G}}}{n}}, \tag{3.2}$$

where g_n is the rule selected by minimizing the empirical risk (2.2) over a class $\mathcal{G}$ with finite Vapnik-Chervonenkis dimension $V_{\mathcal{G}}$. If we are just worried about consistency, this reassures us that nothing is lost as long as we take n large. However, keeping in mind that the distribution of the observations may vary with n, inequality (3.2) does not say much about the asymptotic behavior of the optimal risk $\inf_{g \in \mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\}$. In particular, under Assumption 2.1, it is natural to ask for sufficient conditions under which this risk approaches the asymptotic optimal risk $\inf_{g \in \mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\}$ as n grows. We will be concerned with this question in the rest of the paragraph.

According to Assumption 2.1, the random variable $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ converges in distribution to $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$, or, equivalently, $P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}$ converges weakly to $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$ as $n \to \infty$. Recall that this is equivalent to the statement (see Dudley [9] or Billingsley [1])

$$\int f \, \mathrm{d}P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)} \to \int f \, \mathrm{d}P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$$

for each bounded continuous real function f, or, equivalently,

$$P_{(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)}(A) \to P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}(A)$$

for each continuity set A of $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}$. We recall that the boundary ∂A of A is the closure of A minus its interior, and a continuity set is a Borel set whose boundary satisfies $P_{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})}(\partial A) = 0$.

Before stating our result, we still need some more notation. Given a class $\mathcal{G}$ of symmetric reconstruction rules $g : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$, we endow $\mathcal{G}$ with the pseudo-norm

$$\|g\|_{(\mathbb{X}_1, \mathbb{X}_2)} = \int |g| \, \mathrm{d}P_{(\mathbb{X}_1, \mathbb{X}_2)}.$$

Moreover, for every $g \in \mathcal{G}$, we set

$$G = \big\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : g(x_1, x_2) = 1\big\},$$

and we denote by $\mathbb{G}$ the collection of all such Borel sets G when g varies in $\mathcal{G}$. Finally, for $\delta > 0$, we define the δ-neighborhood of G by

$$G^\delta = \big\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : d\big((x_1, x_2), G\big) < \delta\big\},$$

where d is the distance on the separable Banach space $\mathcal{X}$. Our main result is as follows:

Theorem 3.1 *Let $\mathcal{G}$ be a class of symmetric reconstruction rules. Assume that the following two conditions are satisfied:*

- **(H1)** *The class $\mathcal{G}$ is compact for the pseudo-norm $\|.\|_{(\mathbb{X}_1, \mathbb{X}_2)}$.*
- **(H2)** *One has*
$$\lim_{\delta \to 0} \sup_{G \in \mathbb{G}} P_{(\mathbb{X}_1, \mathbb{X}_2)}\big\{(\partial G)^\delta\big\} = 0.$$

Then, under Assumption 2.1,

$$\inf_{g \in \mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\} \to \inf_{g \in \mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\} \quad \textit{as } n \to \infty.$$

Let us briefly discuss these conditions. **(H1)** requires that the class $\mathcal{G}$ of candidate rules for the graph inference cannot be too large. This assumption essentially means that the parameters which describe the rules in $\mathcal{G}$ take their values in a compact set. Thus, from a practical point of view, this condition does not appear as particularly restrictive. On the other hand, the requirement **(H2)** concerns the uniform limit behavior of the δ-neighborhoods of the sets G in $\mathbb{G}$. According to Billingsley and Topsøe [2], **(H2)** is a necessary and sufficient condition for $\mathbb{G}$ to be a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class, that is

$$\lim_{n\to\infty} \sup_{G\in\mathbb{G}} \left|P_n(G) - P_{(\mathbb{X}_1,\mathbb{X}_2)}(G)\right| = 0 \tag{3.3}$$

holds for every sequence (P_n) that converges weakly to $P_{(\mathbb{X}_1,\mathbb{X}_2)}$. (Of course, even if $\mathbb{G}$ is not a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class, (3.3) will hold for special sequences (P_n) such as $P_n \equiv P_{(\mathbb{X}_1,\mathbb{X}_2)}$).

To be more concrete, suppose that each G in $\mathbb{G}$ is the union of at most $J \geq 1$ measurable, convex and disjoint $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-continuity subsets of $\mathcal{X} \times \mathcal{X}$. In other words,

$$G = \bigcup_{j=1}^{J} K_j,$$

where the K_j's are disjoint measurable convex sets (which possibly depend on G) with $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-null boundary, and J is a fixed positive integer (which does not depend on G). In this case, working out the results of Billingsley and Topsøe [2], one easily deduces that $\mathbb{G}$ is a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class, or, equivalently, that condition **(H2)** is satisfied. For more results on connections between weak and uniform convergence of measures and examples of applications, we refer the reader to Ranga Rao [19] and Topsøe [22].

We summarize the results in the following corollary, which is a consequence of inequality (3.1), Corollary 3.1, and Theorem 3.1:

Corollary 3.2 *Under the conditions of Theorem 3.1, if $\mathcal{G}$ has finite Vapnik-Chervonenkis dimension, we have*

$$\mathbb{E}\mathbf{R}^n(g_n) \to \inf_{g\in\mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1,\mathbb{X}_2) \neq \mathbb{Y}\} \quad \textit{as } n \to \infty,$$

where g_n is the rule which minimizes the empirical reconstruction risk (2.2) over $\mathcal{G}$.

We note finally that the excess reconstruction risk $\mathbb{E}\mathbf{R}^n(g_n) - \mathbf{R}^n_*$ may be decomposed into "estimation" and "approximation" errors as follows:

$$\mathbb{E}\mathbf{R}^n(g_n) - \mathbf{R}^n_* = \left(\mathbb{E}\mathbf{R}^n(g_n) - \inf_{g\in\mathcal{G}} \mathbf{R}^n(g)\right) + \left(\inf_{g\in\mathcal{G}} \mathbf{R}^n(g) - \mathbf{R}^n_*\right).$$

Inequality (3.2) tells that the estimation term vanishes as n grows, whereas Corollary 3.2 ensures that the infimum in the approximation term reaches asymptotically the quantity $\inf_{g\in\mathcal{G}} \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\}$. Here we recall that

$$\mathbf{R}^n_* = \inf_g \mathbf{R}^n(g) = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big\{|2\eta^n(\mathbb{X}^n_1, \mathbb{X}^n_2) - 1|\big\},$$

where the infimum is taken over all measurable reconstruction rules $g : \mathcal{X} \times \mathcal{X} \to \{-1, 1\}$, and $\eta^n(x_1, x_2) = \mathbb{P}\big\{\mathbb{Y}^n = 1|(\mathbb{X}^n_1, \mathbb{X}^n_2) = (x_1, x_2)\big\}$ denotes the posterior probability. Setting

$$\mathbf{R}_* = \frac{1}{2} - \frac{1}{2}\mathbb{E}\big\{|2\eta(\mathbb{X}_1, \mathbb{X}_2) - 1|\big\},$$

with $\eta(x_1, x_2) = \mathbb{P}\big\{\mathbb{Y} = 1|(\mathbb{X}_1, \mathbb{X}_2) = (x_1, x_2)\big\}$, one may ask whether $\mathbf{R}^n_* \to \mathbf{R}_*$ as $n \to \infty$. Sufficient conditions for this consistency to hold are easy to obtain when the sequence $\big(\sigma(\mathbb{X}^n_1, \mathbb{X}^n_2)\big)_n$ is increasing (see Durrett [10], Chapter 4). More general results are difficult to obtain, and beyond the scope of this paper.

We conclude this section with an example to show how to use these results in practice.

Example: *k*-NN rules. In the k-nearest neighbor rule (k-NN) for classification, a majority vote decision is made over the labels of the k nearest neighbors of an unlabeled pattern. Assume, in all generality, that we have at hand a set of ℓ observations $(Z_1, Y_1), \ldots, (Z_\ell, Y_\ell)$ taking their values from $\mathbb{R}^p \times \{-1, 1\}$, and suppose that $\mathbb{R}^p$ is endowed with some norm $\|.\|$. We first reorder the data

$$\big(Z_{(1)}(z), Y_{(1)}(z)\big), \ldots, \big(Z_{(\ell)}(z), Y_{(\ell)}(z)\big)$$

according to increasing distances $\|Z_i - z\|$ of the Z_i's to $z \in \mathbb{R}^p$. In other words, $Z_{(i)}(z)$ is the i-th nearest neighbor of z amongst Z_j, $j = 1, \ldots, \ell$. If distance ties occur, a tie-breaking strategy must be defined. For example, in case of $\|Z_i - z\| = \|Z_j - z\|$, Z_i may be declared closer to z if $i < j$, i.e., the tie-breaking is done by indices. The k-NN rule is then formally defined as

$$g_{\ell,k}(z) = \begin{cases} -1 & \text{if } \sum_{i=1}^k \mathbf{1}_{\{Y_{(i)}(z)=-1\}} \geq \sum_{i=1}^k \mathbf{1}_{\{Y_{(i)}(z)=1\}} \\ 1 & \text{otherwise.} \end{cases}$$

This classification algorithm may be adapted to our graph inference framework in several manners. One natural way is as follows. Given a graph (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$, choose a symmetric and measurable function $\gamma : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^p$, $p \geq 1$, set $Z_{ij}^n = \gamma(X_i^n, X_j^n)$, and consider k-NN classification on the $\mathbb{R}^p \times \{-1, 1\}$-valued transformed data (Z_{ij}^n, Y_{ij}^n), $1 \leq i < j \leq n$. In other words, given a new pair of vertices $(x_1, x_2) \in \mathbb{R}^d \times \mathbb{R}^d$, we set $z = \gamma(x_1, x_2)$, look at the k-nearest neighbors of z amongst (Z_{ij}^n), $1 \leq i < j \leq n$, and decide to draw an edge between x_1 and x_2 according to a majority vote over the labels matching the nearest neighbors of z. Note that γ need not satisfy the axioms of a distance. It is just a function of our choice, which measures the eventual closeness between points in $\mathbb{R}^d$, and which should depend on the problem at hand. A possible choice is for instance the norm $\gamma(X_i^n, X_j^n) = \|X_i^n - X_j^n\|$, but many other alternatives are possible, depending on the context.

We insist on the fact that other strategies than the ordering of the pairs (X_i^n, X_j^n) via the transformation γ may be considered. An interesting alternative approach is to consider the pairs (X_i^n, X_j^n) as vectors in $\mathbb{R}^d \times \mathbb{R}^d$, and to perform k-NN classification directly in this embedding space. The crux here is to define the nearest neighbors of (x_1, x_2) in a symmetric way, i.e., the nearest neighbors of (x_1, x_2) should be the same as those of (x_2, x_1). Clearly, the usual metrics (Euclidean, L_1 or supremum) do not meet this requirement, and tools such as the symmetric Hausdorff distance will certainly be more adapted. We refer to Section 5 for a practical example of application.

Nevertheless, to design an efficient rule, we still need to automatically choose the number k of neighbors. The difficulty here arises from the fact that the nearest neighbor rules are, by definition, data-dependent. To circumvent this problem, we propose to adapt in the graph framework a simple data-splitting device surveyed for example in Devroye [7]. To this aim, we first split the graph artificially into two *independent* subgraphs, one of size $n(n-1)/2$, and one of size $m(m-1)/2$: we call the subgraph based on (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$, the *training graph*, and the subgraph based on (X_i^m, X_j^m, Y_{ij}^m), $n+1 \leq i < j \leq n+m$, the *testing subgraph*. In the sequel, n and the training subgraph are kept fixed, whereas m is allowed to tend to infinity.

To fix ideas, consider as above the transformed data $Z_{ij}^m = \gamma(X_i^m, X_j^m)$ and $Z_{ij}^n = \gamma(X_i^n, X_j^n)$, and let $g_{n,k}$ be the k-NN reconstruction rule based on the training set (Z_{ij}^n, Y_{ij}^n), $1 \leq i < j \leq n$. We select k optimally by minimizing the empirical probability of reconstruction based on the independent testing set. That is, with

$$R_m(g_{n,k}) = \frac{2}{m(m-1)} \sum_{n+1\le i<j\le n+m} \mathbf{1}_{\{g_{n,k}(Z_{ij}^m)\neq Y_{ij}^m\}},$$

we choose

$$\hat{k} \in \arg \min_{1\le k\le n(n-1)/2} R_m(g_{n,k}).$$

Denoting by $\mathcal{G}_n$ the class of all k-NN reconstruction rules based on the training subgraph and setting $g_n \equiv g_{n,\hat{k}}$, this may be rewritten as

$$g_n \in \arg\min_{g\in\mathcal{G}_n} R_m(g).$$

Next, we set $\mathbb{Z}_{12}^m = \gamma(\mathbb{X}_1^m, \mathbb{X}_2^m)$. By Lemma 3.1, the probability of error

$$\begin{aligned}&\mathbf{R}^m(g_n)\\ &= \mathbb{P}\{g_n(\mathbb{Z}_{12}^m) \neq \mathbb{Y}^m | (Z_{ij}^n, Y_{ij}^n), 1\le i<j\le n, (Z_{ij}^m, Y_{ij}^m), n+1\le i<j\le n+m\}\end{aligned}$$

of the selected rule is easily seen to satisfy the inequality

$$\mathbb{E}\mathbf{R}^m(g_n) - \mathbb{E}\Big\{\inf_{g\in\mathcal{G}_n} \mathbf{R}^m(g)\Big\} \le 2\sqrt{\frac{2\log n}{(m-1)^2}} \tag{3.4}$$

since there are only $n(n-1)/2$ possible values for k (see for example Lugosi [15], Section 3.1).

Keeping in mind that n and the learning subgraph are held fixed, and using the fact that the class of reconstruction rules $\mathcal{G}_n$ has finite cardinal $n(n-1)/2$, we deduce that condition **(H1)** is satisfied. Suppose further, for example, that γ is the Euclidean norm. Then, under Assumption 2.1 (that is, $(\mathbb{X}_1^m, \mathbb{X}_2^m, \mathbb{Y}^m)$ converges in law to $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$), the pair $(\mathbb{Z}_{12}^m, \mathbb{Y}^m)$ converges in law toward the pair $(\mathbb{Z}_{12}, \mathbb{Y})$ as $m\to\infty$. Moreover, for each fixed rule $g_{n,k}$ in $\mathcal{G}_n$, the set

$$G_{n,k} = \{z\in\mathbb{R}^+ : g_{n,k}(z) = 1\}$$

is the disjoint union of at most $n(n-1)/2$ intervals. Thus, assuming for example that $\mathbb{X}_1$ is absolutely continuous with respect to the Lebesgue measure, we deduce that condition **(H2)** is verified.

We note finally that if we let m and n tend to infinity, the condition $\log n/m^2 \to 0$ ensures the consistency of the approximation term (3.4). Proving the consistency of the term $\mathbb{E}\{\inf_{g\in\mathcal{G}_n} \mathbf{R}^m(g)\}$ towards the Bayes risk $\mathbf{R}^*$ is more involved, and substantially different arguments are called for (see the discussion page 9).

4 Convex risk minimization

Several successful algorithms for classification, including various versions of boosting and support vector machines, are based on replacing the binary loss function by a convex function and minimizing the corresponding empirical convex risk over a certain class of functions. The main advantage of replacing the empirical probability of misclassification by an appropriate smooth loss functional is to simultaneously avoid overfitting and to deal with computationally feasible algorithms. Recently, significant theoretical advances have been made in understanding the statistical behavior of such methods, see, e.g., Blanchard, Lugosi and Vayatis [3] or Lugosi and Vayatis [16].

In the present section, we wish to briefly extend the principle of convex risk minimization to the graph inference problem considered in the present paper. We start from the arguments presented in Lugosi and Vayatis [16]. To summarize in our context, we consider reconstruction rules of the form

$$g_f(x_1, x_2) = \begin{cases} 1 & \text{if } f(x_1, x_2) > 0 \\ -1 & \text{otherwise,} \end{cases}$$

where $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a symmetric and measurable function. To shorten notation, we will simply write $\mathbf{R}^n(g_f) = \mathbf{R}^n(f)$. Let $\phi : \mathbb{R} \to [0, \infty)$ be a *convex cost function* satisfying $\phi(0) = 1$ and $\phi(x) \geq \mathbf{1}_{\{x \geq 0\}}$. Typical choices of ϕ include the exponential cost function $\phi(x) = e^x$, the "logit" function $\phi(x) = \log_2(1 + e^x)$, or the "hinge" loss $\phi(x) = (1 + x)_+$. Introduce the cost functional associated to ϕ:

$$\mathbf{A}^n(f) = \mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n)$$

and its empirical version based on the observed graph (X_i^n, X_j^n, Y_{ij}^n), $1 \leq i < j \leq n$, defined by

$$A_n(f) = \frac{2}{n(n-1)} \sum_{1 \leq i < j \leq n} \phi(-f(X_i^n, X_j^n)Y_{ij}^n).$$

The reconstruction rule based on convex risk minimization minimizes, over a set $\mathcal{F}$ of candidate symmetric functions $f : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, the empirical cost functional A_n, that is, we choose $f_n \in \arg\min_{f \in \mathcal{F}} A_n(f)$ and assign the corresponding inference rule

$$g_n(x_1, x_2) = \begin{cases} 1 & \text{if } f_n(x_1, x_2) > 0 \\ -1 & \text{otherwise.} \end{cases}$$

(Here we assume implicitly that the minimum exists. More precisely, one may define f_n as any function in $\mathcal{F}$ satisfying $A_n(f_n) \leq \inf_{f\in\mathcal{F}} A_n(f) + 1/n$). Now, setting

$$\mathbf{A}^n(f_n) = \mathbb{E}\big\{\phi(-f_n(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n)|(X_i^n, X_j^n, Y_{ij}^n), 1 \leq i < j \leq n\big\},$$

we may bound, just like in Section 3, the excess convex risk over the class $\mathcal{F}$ as

$$\mathbb{E}\mathbf{A}^n(f_n) - \inf_{f\in\mathcal{F}} \mathbf{A}^n(f) \leq 2\mathbb{E}\Big\{\sup_{f\in\mathcal{F}} \big|A_n(f) - \mathbf{A}^n(f)\big|\Big\}.$$

To bound the right-hand side, we may once again appeal to Lemma 3.1, which gives:

$$\begin{aligned}
&\mathbb{E}\Big\{\sup_{f\in\mathcal{F}} \frac{2}{n(n-1)}\Big|\sum_{1\leq i<j\leq n} \phi(-f(X_i^n, X_j^n)Y_{ij}^n) - \mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n)\Big|\Big\} \\
&\leq \mathbb{E}\Big\{\sup_{f\in\mathcal{F}} \frac{1}{\lfloor n/2\rfloor}\Big|\sum_{i=1}^{\lfloor n/2\rfloor} \phi(-f(X_i^n, X_{i+\lfloor n/2\rfloor}^n)Y_{i\,i+\lfloor n/2\rfloor}^n) - \mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n)\Big|\Big\}.
\end{aligned} \tag{4.1}$$

Now, the right-hand term may be bounded by standard symmetrization and contraction inequalities. As from now, we assume for simplicity that the class $\mathcal{F}$ of reconstruction functions is uniformly bounded, say

$$\sup_{f\in\mathcal{F},(x_1,x_2)\in\mathcal{X}\times\mathcal{X}} \big|f(x_1, x_2)\big| \leq B.$$

We suppose also that the convex cost function ϕ is Lipschitz on the compact set $[-B, B]$ with Lipschitz constant L_B, and that it satisfies $\phi(0) = 1$. Using standard empirical process arguments (Lugosi and Vayatis [16]), we deduce from (4.1) that

$$\mathbb{E}\Big\{\sup_{f\in\mathcal{F}} \big|A_n(f) - \mathbf{A}^n(f)\big|\Big\} \leq 4L_B\mathbb{E}\Big\{\sup_{f\in\mathcal{F}} \frac{1}{\lfloor n/2\rfloor}\Big|\sum_{i=1}^{\lfloor n/2\rfloor} \sigma_i f(X_i^n, X_{i+\lfloor n/2\rfloor}^n)\Big|\Big\},$$

where $\sigma_1, \ldots, \sigma_{\lfloor n/2\rfloor}$ are i.i.d. Rademacher random variables independent of the sample, that is, symmetric sign variables with $\mathbb{P}\{\sigma_i = -1\} = \mathbb{P}\{\sigma_i = 1\} = 1/2$.

Many interesting bounds are available for the Rademacher average of various classes of functions. For example, in many applications such as boosting or bagging, classifiers are combined by weighted voting schemes. This means that the classification rule is obtained using functions f from the class

$$\mathcal{F}_B = \left\{ f(x_1, x_2) = \sum_{j=1}^{N} w_j g_j(x_1, x_2) : N \in \mathbb{N}, \sum_{j=1}^{N} |w_j| \leq B, g_1, \ldots, g_N \in \mathcal{G} \right\}, \tag{4.2}$$

where $\mathcal{G}$ is a family of symmetric reconstruction rules as in Section 3. A classifier of this form may be thought of as one that, upon observing the pair of vertices (x_1, x_2), takes a weighted vote of the classifiers $g_1, \ldots, g_N$ (using the weights $w_1, \ldots, w_N$) and decides according to the weighted majority. In this case, one easily sees that, for some positive universal constant C,

$$\mathbb{E}\left\{ \sup_{f \in \mathcal{F}_B} \frac{1}{\lfloor n/2 \rfloor} \left| \sum_{i=1}^{\lfloor n/2 \rfloor} \sigma_i f(X_i^n, X_{i+\lfloor n/2 \rfloor}^n) \right| \right\} \leq BC \sqrt{\frac{V_{\mathcal{G}}}{n}},$$

where $V_{\mathcal{G}}$ is the Vapnik-Chervonenkis dimension of the base class $\mathcal{G}$. Summarizing, we have thus shown that for the reconstruction rule based on the empirical minimization of the functional A_n over a class $\mathcal{F}_B$ of form (4.2), the excess reconstruction risk satisfies the inequality

$$\mathbb{E}\mathbf{A}^n(f_n) - \inf_{f \in \mathcal{F}_B} \mathbf{A}^n(f) \leq 8 L_B B C \sqrt{\frac{V_{\mathcal{G}}}{n}}. \tag{4.3}$$

Suppose now that Assumption 2.1 is satisfied. Just as in Section 3, we are interested in the risk consistency for a general family $\mathcal{F}$, that is

$$\inf_{f \in \mathcal{F}} \mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n) \to \inf_{f \in \mathcal{F}} \mathbb{E}\phi(-f(\mathbb{X}_1, \mathbb{X}_2)\mathbb{Y}) \quad \text{as } n \to \infty.$$

Denote by $S\big((x_1, x_2), \delta\big)$ the open sphere with center $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$ and radius $\delta > 0$. We define the oscillation of $f \in \mathcal{F}$ on $S\big((x_1, x_2), \delta\big)$ by

$$\omega_f S\big((x_1, x_2), \delta\big) = \sup_{z_1, z_2 \in S((x_1, x_2), \delta)} \Big\{ \big| f(z_1) - f(z_2) \big| \Big\}.$$

Theorem 4.1 *Let $\mathcal{F}$ be a class of uniformly bounded symmetric reconstruction rules. Assume that the following two conditions are satisfied:*

- **(C1)** *The class $\mathcal{F}$ is compact for the pseudo-norm $\|.\|_{(\mathbb{X}_1, \mathbb{X}_2)}$.*
- **(C2)** *One has*

$$\lim_{\delta \to 0} \sup_{f \in \mathcal{F}} P_{(\mathbb{X}_1, \mathbb{X}_2)}\big\{ (x_1, x_2) \in \mathcal{X} \times \mathcal{X} : \omega_f S\big((x_1, x_2), \delta\big) > \varepsilon \big\} = 0.$$

Then, under Assumption 2.1,

$$\inf_{f \in \mathcal{F}} \mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n) \to \inf_{f \in \mathcal{F}} \mathbb{E}\phi(-f(\mathbb{X}_1, \mathbb{X}_2)\mathbb{Y}) \quad \textit{as } n \to \infty.$$

Applying this result along with inequality (4.3) to the class of reconstruction rules (4.2), we immediately obtain:

Corollary 4.1 *Let $\mathcal{F}_B$ be a class of reconstruction rules as in* (4.2)*. Then, under the conditions of Theorem 4.1, if $\mathcal{G}$ has finite Vapnik-Chervonenkis dimension, we have*

$$\mathbb{E}\mathbf{A}^n(f_n) \to \inf_{f\in\mathcal{F}_B} \mathbb{E}\phi(-f(\mathbb{X}_1,\mathbb{X}_2)\mathbb{Y}) \quad \textit{as } n\to\infty.$$

Conditions **(C1)** and **(C2)** are analogous to conditions **(H1)** and **(H2)** of Theorem 3.1. Roughly, **(C1)** means that the set of parameters describing $\mathcal{F}$ is compact, whereas **(C2)** is essentially the functional version of condition **(H2)**. According to Billingsley and Topsøe [2], **(C2)** is a necessary and sufficient condition for $\mathcal{F}$ to be a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class, in the sense that

$$\lim_{n\to\infty}\sup_{f\in\mathcal{F}}\left|\int f\,\mathrm{d}P_n - \int f\,\mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)}\right| = 0$$

holds for every sequence (P_n) that converges weakly to $P_{(\mathbb{X}_1,\mathbb{X}_2)}$. A sufficient condition for a uniformly bounded class $\mathcal{F}$ to be a P-uniformity class for all P is that $\mathcal{F}$ be equicontinuous (Ranga Rao [19], Billingsley and Topsøe [2]). Here, "equicontinuous" means equicontinuous at each point, that is, for every $z\in\mathcal{X}\times\mathcal{X}$ and $\varepsilon>0$ there is a neighborhood U of z such that $\left|f(z)-f(z')\right|\leq\varepsilon$ for all $z'\in U$ and all $f\in\mathcal{F}$. An example of a uniformly bounded equicontinuous family is the set $\mathcal{F}$ of all functions f with $\|f\|_{BL}\leq 1$, where $\|.\|_{BL}$ stands for the bounded Lipschitz norm (Dudley [9], Chapter 11). On the other hand, any finite set of bounded continuous functions is uniformly bounded and equicontinuous, whether or not the functions are Lipschitz.

As an illustration, consider the simple class of functions (4.2) in which the base reconstruction class $\mathcal{G}$ contains all classifiers of the form $g_j(x_1,x_2) = 2\mathbf{1}_{\{\|x_1-x_2\|<a_j\}}-1$, $a_j\in\mathbb{R}^+$. In other words, for each elementary rule g_j, we decide that there is an edge between x_1 and x_2 if the distance between them is smaller than the threshold a_j. Then $V_{\mathcal{G}}<\infty$, and as a general strategy to select the thresholds $a_1,\ldots,a_N$, we can use the empirical risk minimization approach presented in the present paper. In order to apply Corollary 4.1, it is necessary to put some restrictions on $\mathcal{G}$. Suppose for example that the number N of base functions and the set of all possible thresholds in (4.2) are bounded. Suppose also that the random variable $\|\mathbb{X}_1-\mathbb{X}_2\|$ representing the asymptotic distance between the vertices is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^+$, with a bounded density. Then, by a compactness argument it is easy to check that conditions **(C1)** and **(C2)** are satisfied.

5 Experimental results

To illustrate the results of the paper, we implemented several supervised classification algorithms on simulated graphs. The vertices of these graph data sets were independently and uniformly generated points in the bi-dimensional unit square $[0,1] \times [0,1]$. Denoting $X_{ij} = \|X_i - X_j\|$ (the Euclidean norm between X_i and X_j), edges were placed between the pair of vertices $(X_i\,, X_j)$ via the function

$$f(X_{ij}) = \mathbf{1}_{\{X_{ij} < T\}}\, \mathbf{1}_{\{r_1 < 1 - X_{ij}/(20T)\}} + \mathbf{1}_{\{X_{ij} \geq T\}} \mathbf{1}_{\{r_2 < \exp(-X_{ij}/\alpha)\}},$$

where $\alpha > 0$ and $T \in]0\,, \sqrt{2}]$ were fixed (the same for all pairs of vertices) and r_1, r_2 were uniformly generated from the interval $[0,1]$ (differently for each pair of vertices). If $f(X_{ij}) = 1$ we placed an edge, if $f(X_{ij}) = 0$ we did not. Further noise was then included by independently adding or removing the edge between each pair of vertices using a Bernoulli trial with small fixed p. One typical resulting graph is shown in Figure 1. It could model a simple road and town network, where nearby towns are usually reached by a direct road, and faraway towns are reached by passing through several other towns or occasionally by a direct highway.

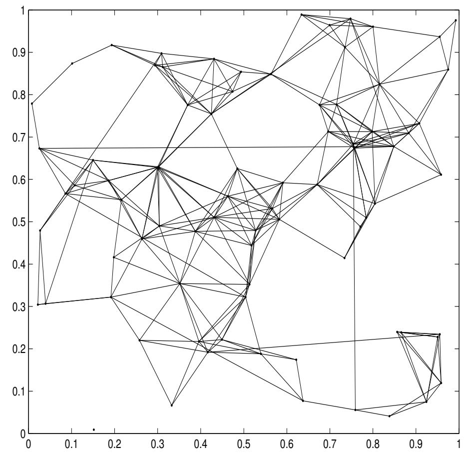

Figure 1: A simulated 2-dimensional graph.

For each of 50 trials, we generated a graph with 110 vertices which was then used for *every* example. The subgraph formed by the first 50 vertices and the edges between these vertices served as the learning set for all of the examples. Two methods, namely k-NN and k-NN Hausdorff, also used the subgraph formed by the next 30 vertices (and internal edges) in the learning step, whereas the SVM methods did not. The final 30 vertices (and internal edges) were used as the test set for all of the methods. The parameters α and T were fixed at 0.01 and 0.4, respectively, and p fixed at 0.02. The results are averaged over the 50 trials and shown in Table 1.

Method	Parameters	Est. error	Sd. of est. error
k-NN	$\overline{k} = 9.28$, $\sigma(k) = 15.41$	0.033	0.009
k-NN Hausdorff	$\overline{k} = 19.80$, $\sigma(k) = 21.28$	0.117	0.025
Linear kernel SVM	$C = 5$	0.100	0.072
Rbf kernel SVM	$C = 5$, $\sigma = 0.1$	0.039	0.010

Table 1: Error results for various reconstruction methods.

For the k-NN and k-NN Hausdorff classifications, $\overline{k}$ is the average number of automatically selected neighbors, and $\sigma(k)$ the standard deviation. The SVM parameters were fixed as shown, before performing the 50 trials.

The k-NN classification (see the example in Section 3) was performed using the Euclidean norm to map each pair of vertices to $\mathbb{R}^+$. A low average estimated true error probability (using the k chosen with the training set to reclassify the test set) was found, due to the fact that the data set contained useful information related to the Euclidean metric. We also performed a k-NN Hausdorff reconstruction, which means finding k-nearest neighbors (of *pairs of vertices*) with respect to the symmetric Hausdorff metric in (for this example) $[0, 1] \times [0, 1]$. We recall the standard definition of the symmetric Hausdorff metric: Let S be a compact metric space (with metric d^*) and let V be the space of nonempty closed and bounded subsets of S. Then the *symmetric Hausdorff metric* is defined on pairs of elements in V as follows: for P, $Q \in V$,

$$d(P, Q) = \max \left\{ \sup_{p \in P} \inf_{q \in Q} d^*(p, q), \sup_{q' \in Q} \inf_{p' \in P} d^*(q', p') \right\}.$$

In terms of the graph data set, we have $P = \{X_i, X_j\}$ and $Q = \{X_k, X_l\}$ as two sets of two vertices of the graph. The hypothesis behind this kind of reconstruction is merely an extension of the usual k-NN hypothesis that an object is likely to share the same label as its nearest neighbors, only here the relevant object is a *pair* of vertices. This classification performs reasonably well on the data set, due to its ability to indirectly use information related to the Euclidean metric.

Finally, we performed support vector machine (SVM) reconstructions on the simulated data set using linear and radial basis function (rbf) kernels. In its general form, a kernel SVM maps the data points (with a function ϕ that may be non-linear) into a (usually) high-dimensional (or even infinitely dimensional) space, in which a hyperplane is chosen to best separate the points into the two classes. If the points are linearly separable in the mapped space, this hyperplane is the solution to the convex optimization problem which gives the hyperplane maximising the *margin* between the closest differently labelled points. If the points are not linearly separable, a *slack* parameter $C > 0$ is added to allow for occasionally misclassified points, resulting in a slightly modified convex optimization problem. The kernel $K(x, y) = \langle \phi(x), \phi(y) \rangle$ is used explicitly in the convex optimization algorithm to measure the distance between the images of the points x and y in the mapped space. Even though the mapped space may be very high or infinitely dimensional, this calculation remains feasible is some cases due to the *kernel trick*, meaning when $K(x, y)$ can be calculated cheaply in the original (often low-dimensional) space. One example is the rbf kernel

$$K(x, y) = \exp\left(- \frac{\|x - y\|^2}{2\sigma^2} \right),$$

for $\sigma > 0$ constant. Other commonly used kernels are the linear kernel and the polynomial kernel. The reader is referred to Schölkopf and Smola's monograph [20] for a comprehensive introduction to SVM and kernels.

For the simulated graph data set, we had to deal with the symmetry constraint (due to the graph being undirected) before performing the SVM algorithms. One way to proceed was to preprocess the data by transforming each pair of vertices into a *point* in $\mathbb{R}^2$ using the map

$$\varphi(X_i, X_j) = \big(|X_i^{(1)} - X_j^{(1)}|, |X_i^{(2)} - X_j^{(2)}| \big).$$

Of course, just as the choice of kernel can represent *a priori* information, so too this map had to be chosen judiciously. We used this transformed data to perform the linear SVM algorithm ($K(x, y) = \langle x, y \rangle$, the usual scalar product) and the rbf kernel SVM algorithm, using the *Spider* Matlab toolbox [21]. Table 1 shows that good results (low estimated true errors) are obtained by both the linear and rbf kernel SVM. In particular, the rbf kernel performs comparably to the k-NN algorithm, the best of the other algorithms.

6 Proofs

6.1 Proof of Lemma 3.1

The result follows from the following chain of inequalities:

$$\begin{aligned}
&\mathbb{E}\Big\{\sup_{\varphi\in\Phi}\frac{2}{n(n-1)}\Big|\sum_{1\le i<j\le n}\varphi(X_i^n,X_j^n,Y_{ij}^n)\Big|\Big\}\\
&\quad=\mathbb{E}\Big\{\sup_{\varphi\in\Phi}\frac{1}{n(n-1)}\Big|\sum_{1\le i\neq j\le n}\varphi(X_i^n,X_j^n,Y_{ij}^n)\Big|\Big\}\\
&\qquad\text{(since each } \varphi \text{ is symmetric in the first two arguments, and } Y_{ij}^n=Y_{ji}^n)\\
&\quad=\mathbb{E}\Big\{\sup_{\varphi\in\Phi}\frac{1}{n!}\Big|\sum_{\pi}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor}\varphi\big(X_{\pi(i)}^n,X_{\pi(i+\lfloor n/2\rfloor)}^n,Y_{\pi(i)\pi(i+\lfloor n/2\rfloor)}^n\big)\Big|\Big\}\\
&\qquad\big(\text{where the sum is taken over all permutations } \pi \text{ of } \{1,\dots,n\}\big)\\
&\quad\le\frac{1}{n!}\sum_{\pi}\mathbb{E}\Big\{\sup_{\varphi\in\Phi}\frac{1}{\lfloor n/2\rfloor}\Big|\sum_{i=1}^{\lfloor n/2\rfloor}\varphi\big(X_{\pi(i)}^n,X_{\pi(i+\lfloor n/2\rfloor)}^n,Y_{\pi(i)\pi(i+\lfloor n/2\rfloor)}^n\big)\Big|\Big\}\\
&\qquad\text{(by the triangle inequality)}\\
&\quad=\mathbb{E}\Big\{\sup_{\varphi\in\Phi}\frac{1}{\lfloor n/2\rfloor}\Big|\sum_{i=1}^{\lfloor n/2\rfloor}\varphi\big(X_i^n,X_{i+\lfloor n/2\rfloor}^n,Y_{ii+\lfloor n/2\rfloor}^n\big)\Big|\Big\}\\
&\qquad\text{(since all } (X_i^n,X_j^n,Y_{ij}^n),\ 1\le i<j\le n, \text{ have the same distribution).}
\end{aligned}$$

6.2 Proof of Theorem 3.1

We start the proof of Theorem 3.1 with two propositions. We denote by $\bar{A}$ (or $(A)^-$) the closure of A (the smallest closed set containing A), int A its interior (the largest open set contained in A), A^c the complement of A, and we recall that $\partial A = \bar{A}\setminus \text{int } A$. Recall also that the symmetric difference $A\triangle B$ between two sets A and B is defined as $(A\setminus B)\cup(B\setminus A)$, where $A\setminus B = A\cap B^c$.

Proposition 6.1 *Let $\mathbb{G}\triangle\mathbb{G}$ be the class of sets $G\triangle G'$, where G and G' vary in $\mathbb{G}$. Then, under condition* **(H2)***,*

$$\lim_{\delta\to 0}\sup_{G\triangle G'\in\mathbb{G}\triangle\mathbb{G}} P_{(\mathbb{X}_1,\mathbb{X}_2)}\Big\{\big(\partial(G\triangle G')\big)^\delta\Big\}=0.$$

Moreover, under Assumption 2.1*,*

$$\sup_{G\triangle G'\in\mathbb{G}\triangle\mathbb{G}} \left|P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(G\triangle G') - P_{(\mathbb{X}_1,\mathbb{X}_2)}(G\triangle G')\right| \to 0 \quad \text{as } n\to\infty.$$

Proof of Proposition 6.1. The first claim of the proposition is a consequence of the following four elementary statements. Let A and B be two sets. Then,

(*i*) $\partial(A\cup B)\subset \partial A\cup\partial B$.

Observe that $(A\cup B)^- = \bar{A}\cup\bar{B}$, and $\text{int }(A\cup B)\supset \text{int }A\cup\text{int }B$. Consequently,

$$\partial(A\cup B)\subset(\bar{A}\cup\bar{B})\setminus(\text{int }A\cup\text{int }B)\subset\partial A\cup\partial B.\blacksquare$$

(*ii*) $\partial(A\cap B)\subset\partial A\cup\partial B$.

Observe that $(A\cap B)^-\subset\bar{A}\cap\bar{B}$, and $\text{int }(A\cap B) = \text{int }A\cap\text{int }B$. Therefore,

$$\partial(A\cap B)\subset(\bar{A}\cap\bar{B})\setminus(\text{int }A\cap\text{int }B)\subset\partial A\cup\partial B.\blacksquare$$

(*iii*) $\partial(A\triangle B)\subset\partial A\cup\partial B$.

Note that, by claim (*i*),

$$\partial(A\triangle B) = \partial\big((A\setminus B)\cup(B\setminus A)\big)\subset\partial(A\setminus B)\cup\partial(B\setminus A).$$

Thus, by (*ii*), we deduce that

$$\partial(A\triangle B)\subset(\partial A\cup\partial B^c)\cup(\partial B\cup\partial A^c),$$

hence the conclusion, since $\partial A^c=\partial A$ and $\partial B^c=\partial B$. $\blacksquare$

(*iv*) $(A\cup B)^\delta\subset A^\delta\cup B^\delta$.

Pick z in $(A\cup B)^\delta$. Since $d\big(z,(A\cup B)\big)<\delta$, there exists z' in $A\cup B$ such that $d(z,z')<\delta$. If $z'\in A$, then $d(z,A)<\delta$, i.e., $z\in A^\delta$. Similarly, if $z'\in B$, then $z\in B^\delta$.$\blacksquare$

To prove the first statement of Proposition 6.1, note that by (*iii*) and (*iv*),

$$\big(\partial(G\triangle G')\big)^{\delta} \subset (\partial G \cup \partial G')^{\delta} \subset (\partial G)^{\delta} \cup (\partial G')^{\delta}.$$

Therefore, for $\delta > 0$, by the union bound,

$$\sup_{G\triangle G' \in \mathbb{G}\triangle\mathbb{G}} P_{(\mathbb{X}_1,\mathbb{X}_2)}\Big\{\big(\partial(G\triangle G')\big)^{\delta}\Big\} \leq 2 \sup_{G\in\mathbb{G}} P_{(\mathbb{X}_1,\mathbb{X}_2)}\big\{(\partial G)^{\delta}\big\},$$

hence the conclusion under condition **(H2)**.

By virtue of Theorem 3 in Billingsley and Topsøe [2], we have just shown that the class $\mathbb{G}\triangle\mathbb{G}$ is a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-continuity class. This proves, under Assumption 2.1, the second statement of the proposition. ■

Proposition 6.2 *Assume that condition* **(H2)** *holds. Then, for each fixed g in $\mathcal{G}$, under Assumption* 2.1,

$$\mathbb{P}\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\} \to \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\} \quad \text{as } n \to \infty.$$

Proof of Proposition 6.2. Recall that we denote by G the set

$$G = \big\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : g(x_1, x_2) = 1\big\},$$

and observe that, under **(H2)**, we have

$$P_{(\mathbb{X}_1,\mathbb{X}_2)}\{\partial G\} = 0. \tag{6.1}$$

Let us introduce D_g, the subset of $\mathcal{X} \times \mathcal{X} \times \{-1, 1\}$ defined by

$$D_g = \big\{(x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \{-1, 1\} : g(x_1, x_2) = y\big\}.$$

We note that ∂D_g precisely equals the set of discontinuities of the function $(x_1, x_2, y) \mapsto \mathbf{1}_{\{g(x_1,x_2)\neq y\}}$. Also, clearly,

$$\begin{aligned} D_g = &\big\{(x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \{-1\} : g(x_1, x_2) = -1\big\} \\ &\cup \big\{(x_1, x_2, y) \in \mathcal{X} \times \mathcal{X} \times \{1\} : g(x_1, x_2) = 1\big\}, \end{aligned}$$

and, therefore,

$$\begin{aligned} \partial D_g &= \big(\partial G^c \times \{-1\}\big) \cup \big(\partial G \times \{1\}\big) \\ &= \big(\partial G \times \{-1\}\big) \cup \big(\partial G \times \{1\}\big) \qquad (6.2) \\ &\quad \text{(since } \partial G^c = \partial G). \end{aligned}$$

Consequently,

$$
\begin{aligned}
&\mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y}) \in \partial D_g\big\} \\
&\quad = \mathbb{P}\big\{(X_1, X_2, -1) \in \partial D_g, Y = -1\big\} + \mathbb{P}\big\{(X_1, X_2, 1) \in \partial D_g, Y = 1\big\} \\
&\quad \le \mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2) \in \partial G\big\} + \mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2) \in \partial G\big\} \\
&\qquad \text{(by equality (6.2))} \\
&\quad = 2P_{(\mathbb{X}_1, \mathbb{X}_2)}\{\partial G\} \\
&\quad = 0 \qquad (6.3) \\
&\qquad \text{(by equality (6.1))}.
\end{aligned}
$$

Since $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ converges in distribution to $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$ as $n \to \infty$, it follows from (6.3) that $\mathbf{1}_{\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\}}$ converges in distribution to $\mathbf{1}_{\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\}}$ (see for example Billingsley [1], Chapter 5, Section 25, Corollary 1). Taking the expectation, this implies (Billingsley [1], Chapter 5, Theorem 25.12) that

$$\mathbb{P}\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\} \to \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\} \quad \text{as } n \to \infty,$$

as desired. ■

We are now in a position to prove Theorem 3.1. To lighten notation, for each fixed $g \in \mathcal{G}$, we set

$$l_n(g) = \mathbb{P}\big\{g(\mathbb{X}_1^n, \mathbb{X}_2^n) \neq \mathbb{Y}^n\big\}.$$

By condition **(H1)**, $\mathcal{G}$ is compact for the pseudo-norm $\|.\|_{(\mathbb{X}_1, \mathbb{X}_2)}$. Moreover, by Proposition 6.2 above,

$$l_n(g) \to \mathbb{P}\{g(\mathbb{X}_1, \mathbb{X}_2) \neq \mathbb{Y}\} \quad \text{as } n \to \infty.$$

Therefore, to prove Theorem 3.1, it suffices to show that the sequence $\big(l_n(.)\big)_n$ is equicontinuous for all n large enough. (That is, for every $\varepsilon > 0$ and all n large enough, for every $g \in \mathcal{G}$, there is a neighborhood U of g such that $\big\|l_n(g) - l_n(g')\big\|_{(\mathbb{X}_1, \mathbb{X}_2)} \le \varepsilon$ for all $g' \in \mathcal{G}$). To show this, let $\varepsilon > 0$ and g and g' be two reconstruction rules in $\mathcal{G}$ such that $\|g - g'\|_{(\mathbb{X}_1, \mathbb{X}_2)} \le \varepsilon$. Denote by G and G' the associated reconstruction sets, that is

$$G = \big\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : g(x_1, x_2) = 1\big\}$$

and

$$G' = \big\{(x_1, x_2) \in \mathcal{X} \times \mathcal{X} : g'(x_1, x_2) = 1\big\}.$$

We have

$$
\begin{aligned}
&\big|l_n(g) - l_n(g')\big| \\
&\quad = \int_{G\triangle G'} |2\eta^n(x_1,x_2) - 1| \, \mathrm{d}P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(x_1,x_2) \\
&\qquad \text{(see e.g. Devroye, Györfi and Lugosi [8], Chapter 2, Theorem 2.2)} \\
&\quad \le P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(G\triangle G') \\
&\qquad \text{(using the fact } \eta^n(x_1,x_2) \le 1\text{)} \\
&\quad \le P_{(\mathbb{X}_1,\mathbb{X}_2)}(G\triangle G') + \sup_{G\triangle G' \in \mathbb{G}\triangle\mathbb{G}} \big|P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(G\triangle G') - P_{(\mathbb{X}_1,\mathbb{X}_2)}(G\triangle G')\big| \\
&\quad = \frac{1}{2}\int |g-g'| \, \mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)} + \sup_{G\triangle G' \in \mathbb{G}\triangle\mathbb{G}} \big|P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(G\triangle G') - P_{(\mathbb{X}_1,\mathbb{X}_2)}(G\triangle G')\big| \\
&\qquad \text{(since } |g-g'| \equiv 2 \text{ on } G\triangle G' \text{ and } |g-g'| = 0 \text{ elsewhere)} \\
&\quad = \frac{1}{2}\|g-g'\|_{(\mathbb{X}_1,\mathbb{X}_2)} + \sup_{G\triangle G' \in \mathbb{G}\triangle\mathbb{G}} \big|P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)}(G\triangle G') - P_{(\mathbb{X}_1,\mathbb{X}_2)}(G\triangle G')\big| \\
&\quad \le \varepsilon/2 + \varepsilon/2 \\
&\qquad \text{(for all } n \text{ large enough, by Proposition 6.1)} \\
&\quad = \varepsilon.
\end{aligned}
$$

■

6.3 Proof of Theorem 4.1

We start by stating two propositions.

Proposition 6.3 *Assume that condition* **(C2)** *holds. Then, under Assumption* 2.1*,*

$$\sup_{(f,f')\in\mathcal{F}^2} \left| \int |f-f'| \, \mathrm{d}P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)} - \int |f-f'| \, \mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)} \right| \to 0 \quad \text{as } n\to\infty.$$

Proof of Proposition 6.3. According to Theorem 1 in Billingsley and Topsøe [2], condition **(C2)** ensures that the bounded family $\mathcal{F}$ is a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class. It immediately follows that the family $\{f+f' : (f,f')\in\mathcal{F}^2\}$ is also a $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-uniformity class, hence the conclusion. ■

Proposition 6.4 *Assume that condition* **(C2)** *holds. Then, for each fixed f in $\mathcal{F}$, under Assumption* 2.1*,*

$$\mathbb{E}\phi(-f(\mathbb{X}_1^n,\mathbb{X}_2^n)\mathbb{Y}^n) \to \mathbb{E}\phi(-f(\mathbb{X}_1,\mathbb{X}_2)\mathbb{Y}) \quad \text{as } n\to\infty.$$

Proof of Proposition 6.4. Denote by D_f the set of discontinuities of the function $(x_1, x_2, y) \mapsto \phi(-f(x_1, x_2)y)$, and observe that

$$D_f = \big(\tilde{D}_f \times \{-1\}\big) \cup \big(\tilde{D}_f \times \{1\}\big), \tag{6.4}$$

where $\tilde{D}_f$ represents the discontinuity points of the mapping $(x_1, x_2) \mapsto \phi\big(-f(x_1, x_2)\big)$. It is easily seen, under condition **(C2)**, that the function f is continuous except on a set of $P_{(\mathbb{X}_1,\mathbb{X}_2)}$-measure 0. Therefore, since ϕ is continuous on the range of f,

$$P_{(\mathbb{X}_1,\mathbb{X}_2)}(\tilde{D}_f) = 0. \tag{6.5}$$

Thus

$$\begin{aligned}
&\mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y}) \in D_f\big\} \\
&\quad = \mathbb{P}\big\{(X_1, X_2, -1) \in D_f, Y = -1\big\} + \mathbb{P}\big\{(X_1, X_2, 1) \in D_f, Y = 1\big\} \\
&\quad \leq \mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2) \in \tilde{D}_f\big\} + \mathbb{P}\big\{(\mathbb{X}_1, \mathbb{X}_2) \in \tilde{D}_f\big\} \\
&\qquad \text{(by equality (6.4))} \\
&\quad = 0 \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (6.6) \\
&\qquad \text{(by equality (6.5))}.
\end{aligned}$$

By Assumption 2.1, $(\mathbb{X}_1^n, \mathbb{X}_2^n, \mathbb{Y}^n)$ converges in distribution to $(\mathbb{X}_1, \mathbb{X}_2, \mathbb{Y})$ as $n \to \infty$. Therefore, using (6.6), we deduce that $\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n)$ converges in distribution to $\phi(-f(\mathbb{X}_1, \mathbb{X}_2)\mathbb{Y})$ (Billingsley [1], Chapter 5, Section 25, Corollary 1). Observing finally that the function $(x_1, x_2, y) \mapsto \phi(-f(x_1, x_2)y)$ is bounded, we conclude (Billingsley [1], Chapter 5, Theorem 25.12) that

$$\mathbb{E}\phi(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n) \to \mathbb{E}\phi(-f(\mathbb{X}_1, \mathbb{X}_2)\mathbb{Y}) \quad \text{as } n \to \infty,$$

as desired. ■

We are now ready to prove Theorem 4.1. Just as in the proof of Theorem 3.1, for each fixed $f \in \mathcal{F}$, we set

$$l_n(f) = \mathbb{E}\phi\big(-f(\mathbb{X}_1^n, \mathbb{X}_2^n)\mathbb{Y}^n\big).$$

Having disposed of the preliminary Proposition 6.4, and invoking the compactness of $\mathcal{F}$ for the pseudo-norm $\|.\|_{(\mathbb{X}_1,\mathbb{X}_2)}$, we will have established the theorem if we prove that the sequence $\big(l_n(.)\big)_n$ is equicontinuous for all n large enough.

Fix $\varepsilon > 0$, and let f and f' be two reconstruction functions in $\mathcal{F}$ such that $\|f-f'\|_{(\mathbb{X}_1,\mathbb{X}_2)} \leq \varepsilon/(2L_B)$, where L_B is the Lipschitz constant of ϕ on $[-B, B]$. Clearly, for all $(x_1, x_2) \in \mathcal{X} \times \mathcal{X}$,

$$\big|\phi\big(-f(x_1,x_2)\big) - \phi\big(-f'(x_1,x_2)\big)\big| \leq L_B \big|f(x_1,x_2) - f'(x_1,x_2)\big|.$$

Consequently,

$$\begin{aligned}
&\big|l_n(f) - l_n(f')\big| \\
&\quad = \big|\mathbb{E}\phi(-f(\mathbb{X}_1^n,\mathbb{X}_2^n)\mathbb{Y}^n) - \mathbb{E}\phi(-f'(\mathbb{X}_1^n,\mathbb{X}_2^n)\mathbb{Y}^n)\big| \\
&\quad \leq \mathbb{E}\big|\phi(-f(\mathbb{X}_1^n,\mathbb{X}_2^n)\mathbb{Y}^n) - \phi(-f'(\mathbb{X}_1^n,\mathbb{X}_2^n)\mathbb{Y}^n)\big| \\
&\quad \leq L_B \mathbb{E}\big|f(\mathbb{X}_1^n,\mathbb{X}_2^n) - f'(\mathbb{X}_1^n,\mathbb{X}_2^n)\big| \\
&\quad = L_B \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)} \\
&\quad \leq L_B \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)} \\
&\qquad + L_B \sup_{(f,f')\in\mathcal{F}^2} \left| \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)} - \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)} \right| \\
&\quad = L_B \|f - f'\|_{(\mathbb{X}_1,\mathbb{X}_2)} \\
&\qquad + L_B \sup_{(f,f')\in\mathcal{F}^2} \left| \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1^n,\mathbb{X}_2^n)} - \int |f - f'|\,\mathrm{d}P_{(\mathbb{X}_1,\mathbb{X}_2)} \right| \\
&\quad \leq \varepsilon/2 + \varepsilon/2 \\
&\qquad \text{(for all } n \text{ large enough, by Proposition 6.3)} \\
&\quad = \varepsilon.
\end{aligned}$$

■

Acknowledgments. The authors thank an Associate Editor and a referee for many valuable comments and suggestions.

References

[1] Billingsley, P. (1995). *Probability and Measure, Third Edition*, John Wiley & Sons, New York.

[2] Billingsley, P. and Topsøe, F. (1967). Uniformity in weak convergence, *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, Vol. 7, pp. 1–16.

[3] Blanchard, G., Lugosi, G. and Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers, *Journal of Machine Learning Research*, Vol. 4, pp. 861–894.

[4] Bollobás, B. (2001). *Random Graphs, Second Edition*, Cambridge University Press, Cambridge.

[5] Boucheron, S., Bousquet, O. and Lugosi, G. (2005). Theory of classification: A survey of some recent advances, *ESAIM: Probability and Statistics*, Vol. 9, pp. 323–375.

[6] Clémençon, S., Lugosi, G. and Vayatis, N. (2006). Ranking and empirical risk minimization of U-statistics, *Technical report, University Paris VI*, http://www.proba.jussieu.fr/pageperso/vayatis/publication.html .

[7] Devroye, L. (1988). Automatic pattern recognition: A study of the probability of error, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 10, pp. 530–543.

[8] Devroye, L., Györfi, L. and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*, Springer–Verlag, New York.

[9] Dudley, R.M. (1989). *Real Analysis and Probability*, Chapman and Hall, New York.

[10] Durrett, R. (1996). *Probability: Theory and Examples, Second Edition*, Duxbury Press, Belmont.

[11] Friedman, N., Linial, M., Nachman, I. and Pe'er, D. (2000). Using Bayesian networks to analyse expression data, *Journal of Computational Biology*, Vol. 7, pp. 601–620.

[12] Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables, *Journal of the American Statistical Association*, Vol. 58, pp. 13–30.

[13] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F. and Gerstein, M. (2003). A Bayesian network approach for predicting protein-protein interactions from genomic data, *Science*, Vol. 302, pp. 449–453.

[14] Kanehisa, M. (2001). Prediction of higher order functional networks from genomic data, *Pharmacogenomics*, Vol. 2, pp. 373–385.

[15] Lugosi, G. (2002). Pattern classification and learning theory, in *Principles of Nonparametric Learning*, ed. L. Györfi, Springer, Wien, pp. 1–56.

[16] Lugosi, G. and Vayatis, N. (2004). On the Bayes-risk consistency of boosting methods (with discussion), *The Annals of Statistics*, Vol. 32, pp. 30–55.

[17] Mattick, J.S. and Gagen, M.J. (2005). Accelerating networks, *Science*, Vol. 307, pp. 856–858.

[18] de la Peña, V.H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*, Springer–Verlag, New York.

[19] Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications, *The Annals of Mathematical Statistics*, Vol. 33, pp. 659–680.

[20] Schölkopf, B. and Smola, A.J. (2002). *Learning with Kernels*, MIT Press, Massachusetts.

[21] *The Spider for Matlab.* http://www.kyb.tuebingen.mpg.de/bs/people/spider .

[22] Topsøe, F. (1967). On the connection between P-continuity and P-uniformity in weak convergence, *Theory of Probability and its Applications*, Vol. 12, pp. 279–288.

[23] Vapnik, V.N. and Chervonenkis, A.Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities, *Theory of Probability and its Applications*, Vol. 16, pp. 264–280.

[24] Vert, J.P. and Yamanishi, Y. (2005). Supervised graph inference, in *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, ed. Y. Weiss and L. Bottou, MIT Press, Cambridge, pp. 1433–1440.