Machine Learning (2024) 113:3929–3959
https://doi.org/10.1007/s10994-024-06536-9

Generalization bounds for learning under graph-dependence: a survey

Rui-Ray Zhang[1,2] · **Massih-Reza Amini**[3]

Received: 18 May 2022 / Revised: 29 January 2024 / Accepted: 5 March 2024 /
Published online: 3 April 2024

Abstract
Traditional statistical learning theory relies on the assumption that data are identically and independently distributed (i.i.d.). However, this assumption often does not hold in many real-life applications. In this survey, we explore learning scenarios where examples are dependent and their dependence relationship is described by a *dependency graph*, a commonly utilized model in probability and combinatorics. We collect various graph-dependent concentration bounds, which are then used to derive Rademacher complexity and stability generalization bounds for learning from graph-dependent data. We illustrate this paradigm through practical learning tasks and provide some research directions for future work. To our knowledge, this survey is the first of this kind on this subject.

Keywords Generalization bounds · Dependency graphs · Uniform stability · Rademacher complexity · Bipartite ranking

1 Introduction

The central assumption in machine learning is that observations are independently and identically distributed (i.i.d.) with respect to a fixed yet unknown probability distribution. Under this assumption, generalization error bounds, shedding light on the learnability of models or conducting in the design of advanced algorithms (Boser et al., 1992), have been proposed. However, in many real applications, the data collected can be dependent, and therefore the i.i.d. assumption does not hold. There have been extensive discussions in the

Editor: Aryeh Kontorovich.

✉ Rui-Ray Zhang
rui.zhang@bse.eu

Massih-Reza Amini
massih-reza.amini@imag.fr

[1] Barcelona School of Economics, 08005 Barcelona, Catalonia, Spain

[2] School of Mathematics, Monash University, Clayton, VIC 3168, Australia

[3] LIG/CNRS, University Grenoble Alpes, 38041 CEDEX 9 Grenoble, France

community on why and how the data are dependent (Amini & Usunier, 2015; Dehling & Philipp, 2002).

Learning with interdependent data Establishing generalization theories under dependent settings have received a surge of interest in recent years (Kuznetsov & Mohri, 2017; Mohri & Rostamizadeh, 2008, 2009; Ralaivola et al., 2010). A major line of research in this direction models the data dependencies by various types of mixing models, such as α-mixing (Rosenblatt, 1956), β-mixing (Volkonskii & Rozanov, 1959), ϕ-mixing (Ibragimov, 1962), and η-mixing (Kontorovich, 2007), and so on. Mixing models have been used in statistical learning theory to establish generalization error bounds based on Rademacher complexity (Kuznetsov & Mohri, 2017; Mohri & Rostamizadeh, 2009, 2010) or algorithmic stability (He et al., 2016; Mohri & Rostamizadeh, 2008, 2010) via concentration results (Kontorovich & Ramanan, 2008) or independent block technique (Yu, 1994). In these models, the mixing coefficients quantitatively measure the dependencies among data. Another line of work, referred to as decoupling, studies the behavior of complex systems by decomposing a set of dependent random variables into sets of independent variables and a set of dependent variables with vanishing moments (Peña & Giné, 1999). A random variable with vanishing moments has a property that its expected value converges to zero as the number of terms increases. This technique of decoupling has been successfully applied in many areas of mathematics, statistics, and engineering.

Dependency graphs Although the results based upon the mixing model and decoupling with vanishing moments are fruitful, they face difficulties in practical applications, as it is usually difficult to determine or estimate the quantitative dependencies among data points (such as the mixing coefficients or the vanishing moments) unless under some restrictive assumptions. On the other hand, determining whether two data are dependent or exhibit a suitable dependency structure is often much easier in practice. Thus in this paper, we focus on such a qualitative dependent setting. We use graphs as a natural tool to describe the dependencies among data and establish generalization theory under such graph-dependence. The dependency graph model we use has been widely utilized in many other fields, in particular, in probability theory and statistics, where it is used to prove normal or Poisson approximation using Stein's approach, cumulants, and so on (see, for example, Janson, 1988, 1990). It is also heavily used in probabilistic combinatorics and statistical physics, such as Lovász local lemma (Erdős & Lovász, 1975), Janson's inequality (Janson et al., 1988), along with many others.

Rademacher complexity We collect various concentration bounds under graph-dependence and utilize them to derive Rademacher and stability generalization bounds for learning from dependent data. The basic tool used to establish generalization theory is concentration inequalities. Standard concentration results for the i.i.d. case no longer apply for dependently distributed data, making the study a challenging task. Janson (2004) extended Hoeffding's inequality to the sum of dependent random variables. This result bounds the probability that the summation of graph-dependent random variables deviates from its expected value, in terms of the fractional chromatic number of the dependency graph. Our first approach uses a similar idea, by dividing graph-dependent variables into sets of independent ones, we establish concentration bounds based on fractional colorings, and generalization bounds via fractional Rademacher complexity.

Algorithmic stability PAC-Bayes bounds for classification with non-i.i.d. data have also been obtained based on fractional colorings of graphs in Ralaivola et al. (2010). These results also hold for specific learning settings such as ranking and learning from stationary β-mixing distributions. Ralaivola and Amini (2015) established new concentration inequalities for fractionally sub-additive and fractionally self-bounding functions

of dependent variables. Though fundamental and elegant, the above generalization bounds are algorithm-independent. They consider the complexity of the hypothesis space and data distribution, but do not involve specific learning algorithms. To derive better generalization bounds, there is growing interest in developing algorithm-dependent generalization theories. This line of research heavily relies on the notion of algorithmic stability, which exhibits a key advantage, that is, they are tailored to specific learning algorithms, exploiting their particular properties. Our second approach utilizes algorithmic stability to establish generalization bounds. Note that even under the i.i.d. assumption, Hoeffding-type concentration inequalities, which bound the deviation of sample average from expectation, are not strong enough to prove stability-based generalization. On the contrary, McDiarmid's inequality characterizes the concentration of general Lipschitz functions of i.i.d. random variables, hence is used as the key tool for proving the stability bounds. Therefore, to build algorithmic stability theory for non-i.i.d. samples, we start with McDiarmid-type concentration bounds for graph-dependent random variables.

Table 1 lists some generalization results using Rademacher complexity and algorithmic stability for i.i.d., mixing, and graph-dependent settings, respectively.

Paper organization In this survey, we begin with introducing different McDiarmid-type concentration inequalities for functions of graph-dependent random variables. Then we utilize these concentration bounds to provide upper bounds on generalization error for learning from graph-dependent data using Rademacher complexity and algorithm stability. In the reminder, Sect. 2 introduces notation and the framework. Section 3 establishes fractional Rademacher complexity and algorithmic stability bounds. Section 4 shows how the presented framework can be utilized to derive generalization bounds for learning from graph-dependent data in a variety of practical scenarios, including learning-to-rank, multi-class classification problems, and learning from m-dependent data. We finally conclude this work in Sect. 5 and provide some perspective and future work.

2 Notation and framework

Throughout this paper, for all positive integer n, let $[n]$ denote the integer set $\{1, 2, \dots, n\}$. Given two integers $i < j$, let $[i, j]$ denote the integer set $\{i, i+1, \dots, j-1, j\}$. Let Ω_i be a Polish space for every $i \in [n]$, $\mathbf{\Omega} = \prod_{i\in[n]} \Omega_i = \Omega_1 \times \dots \times \Omega_n$ be the product space, $\mathbb{R}$ be the set of real numbers, and $\mathbb{R}_+$ be the set of non-negative real numbers. Let $\|\cdot\|_p$ denote the standard ℓ_p-norm of a vector. We use uppercase letters for random variables, lowercase letters for their realizations, and bold letters for vectors.

Table 1 Rademacher complexity and stability generalization bounds for i.i.d., mixing, and graph-dependent settings

	Rademacher bounds	Stability bounds
i.i.d	Bartlett and Mendelson (2002)	Bousquet and Elisseeff (2002)
Mixing conditions	Mohri and Rostamizadeh (2009)	Mohri and Rostamizadeh (2008)
Graph-dependence	Theorem 3.5 (Amini & Usunier, 2015)	Theorem 3.12 (Zhang et al., 2019)

2.1 Graph-theoretic notation

We use the standard graph-theoretic notation. All graphs considered are finite, undirected, and simple (no loops or multiple edges). A graph $G = (V, E)$ consists of a set of vertices V, some of which are connected by edges in E. Given a graph G, let $V(G)$ be the vertex set and $E(G)$ be the edge set. The edge connecting a pair of distinct vertices u, v is denoted by $\{u, v\}$, which is assumed to be unordered. The number of edges incident on a vertex is the degree of the vertex; and we use $\Delta(G)$ to denote the maximum degree of graph G.

2.1.1 Graph covering and partitioning

Formally, given a graph G, we introduce the following definitions.

(a1) A family $\{S_k\}_k$ of subsets of $V(G)$ is a *vertex cover* of G if $\bigcup S_k = V(G)$.

(a2) A vertex cover $\{S_k\}_k$ of G is a *vertex partition* of G if every vertex of G is in exactly one element of $\{S_k\}_k$.

(a3) A family $\{(S_k, w_k)\}_k$ of pairs (S_k, w_k), where $S_k \subseteq V(G)$ and $w_k \in [0, 1]$ is a *fractional vertex cover* of G if $\{S_k\}_k$ is a vertex cover of G, and $\sum_{k: v \in S_k} w_k = 1$ for every $v \in V(G)$.

(a4) An independent set of G is a set of vertices of G, no two of which are adjacent in G. Let $\mathcal{I}(G)$ denote the set of all independent sets of graph G.

(a5) A fractional independent vertex cover $\{(I_k, w_k)\}_k$ of G is a fractional vertex cover such that $I_k \in \mathcal{I}(G)$ for every k.

(a6) A fractional coloring of a graph G is a mapping g from $\mathcal{I}(G)$ to [0, 1] such that $\sum_{I \in \mathcal{I}(G): v \in I} g(I) \geqslant 1$ for every vertex $v \in V(G)$. The fractional chromatic number $\chi_f(G)$ of G is the minimum of the value $\sum_{I \in \mathcal{I}(G)} g(I)$ over fractional colorings of G. See Fig. 1 for an example. Note that the fractional chromatic number $\chi_f(G)$ of graph G is the minimum of $\sum_k w_k$ over all fractional independent vertex covers $\{(I_k, w_k)\}_k$ of G (see, for example, Janson, 2004).

(a7) Let H be a graph and $\{H_x \subseteq V(G)\}_{x \in V(H)}$ be a set of subsets of $V(G)$ indexed by the vertices of H. Each set H_x is called a 'bag'. The pair $(H, \{H_x\}_{x \in V(H)})$ is an H-partition of G if:

(i) $\{H_x\}_{x \in V(H)}$ is a vertex partition of G.

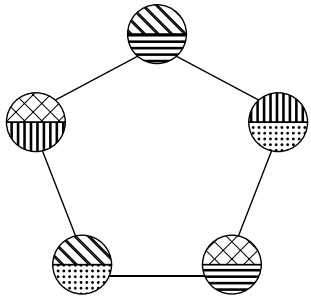


Fig. 1 A fractional coloring of a cycle graph C_5 of length 5 with patterns indicating different colors. The set of pairs $\{(\{i, (i+3) (\text{mod } 5)\}, 1/2)\}_{1 \leqslant i \leqslant 5}$ is a fractional vertex cover with the fractional chromatic number 5/2

(ii) Distinct u and v are adjacent in H if and only if there is an edge of G with one endpoint in H_u and the other endpoint in H_v.

In graph theory, a *vertex identification* (also called vertex contraction) is to contract a pair of vertices u and v of a graph and produces a graph in which the two vertices u and v are replaced with a single vertex t such that t is adjacent to the union of the vertices to which u and v were originally adjacent. Note that in vertex contraction, it does not matter if u and v are connected by an edge; if they are, the edge is simply removed upon contraction, this special case of vertex identification called *edge contraction*.

Informally speaking, an H-partition of graph G is obtained from a proper partition of $V(G)$ by identifying the vertices in each part, deleting loops, and replacing parallel edges with a single edge. H is also called the *quotient graph* of the graph G. For brevity, we say H is a partition of G. For more about partitions of graphs, see, for example, (Wood, 2009).

(a8) A tree is a connected, acyclic graph, and a forest is a disjoint union of trees. For a given forest F, we denote the set of (vertex sets of) disjoint trees in forest F as $\mathcal{T}(F)$.

(a9) If forest F is a partition of graph G, then the pair $(F, \{F_x \subseteq V(G)\}_{x\in V(F)})$ is a *tree-partition* of G. The set of all tree-partitions of graph G is denoted by $\mathsf{TP}(G)$. See Fig. 2 for an example.

Tree-partitions were independently introduced by Seese (1985) and Halin (1991), and have since been widely investigated (Wood, 2009). Essentially, a tree-partition of a graph is a proper partition of its vertex set into 'bags', such that identifying the vertices in each bag produces a forest.

2.2 Probabilistic tools

Concentration inequalities are fundamental tools in statistical learning theory. They bound the deviation of a function of random variables from some value that is usually the expectation. Among the most powerful ones is McDiarmid's inequality (McDiarmid, 1989), which establishes sharp concentration for multivariate functions that do not depend too much on any individual coordinate, specifically, when the function satisfies $\mathbf{c}$ -Lipschitz condition for a weighted hamming distance (bounded differences condition).

Let $\mathbf{1}_{\{A\}}$ denote the indicator function for any event A, that is, $\mathbf{1}_{\{A\}} = 1$ if A occurs, otherwise, $\mathbf{1}_{\{A\}} = 0$. We first introduce the definition of a Lipschitz function.

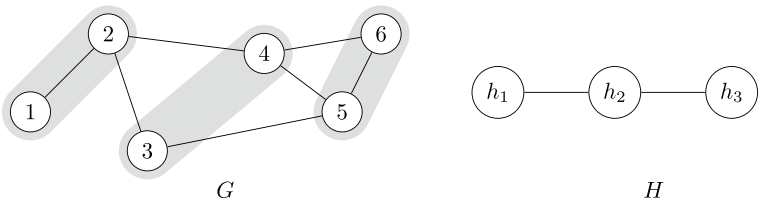


Fig. 2 A tree-partition of graph G is $(H, \{\{1,2\},\{3,4\},\{5,6\}\})$, where H is a path on vertices $\{h_1, h_2, h_3\}$, which correspond to vertex sets $\{1,2\}$, $\{3,4\}$, and $\{5,6\}$ respectively

Definition 2.1 (**c**-Lipschitz) Given a vector $\mathbf{c} = (c_1, \ldots, c_n) \in \mathbb{R}_+^n$, a function $f : \mathbf{\Omega} \to \mathbb{R}$ is **c**-Lipschitz if for all $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{x}' = (x_1', \ldots, x_n') \in \mathbf{\Omega}$, we have

$$|f(\mathbf{x}) - f(\mathbf{x}')| \leqslant \sum_{i=1}^{n} c_i \mathbf{1}_{\{x_i \neq x_i'\}}, \tag{2.1}$$

where c_i is the i-th Lipschitz coefficient of f (with respect to the Hamming metric).

McDiarmid's inequality is based on the following bound on the moment-generating function.

Lemma 2.2 (McDiarmid, 1989) *Let* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a vector of independent random variables taking values in* $\mathbf{\Omega}$ *and* $f : \mathbf{\Omega} \to \mathbb{R}$ *be* **c**-*Lipschitz. Then for any* $s > 0$,

$$\mathbb{E}\left[e^{s(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}))}\right] \leqslant \exp\left(\frac{s^2}{8}\|\mathbf{c}\|_2^2\right).$$

We can now state the following McDiarmid's inequality, which constitutes one of the pillars of our results. It states that a Lipschitz function of independent random variables concentrates around its expectation.

Theorem 2.3 (McDiarmid's inequality 1989) *Let* $f : \mathbf{\Omega} \to \mathbb{R}$ *be* **c**-*Lipschitz and* $\mathbf{X} = (X_1, \ldots, X_n)$ *be a vector of independent random variables that takes values in* $\mathbf{\Omega}$. *Then for every* $t > 0$,

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\|\mathbf{c}\|_2^2}\right). \tag{2.2}$$

In the following, we extend McDiarmid's inequality to the graph-dependent case, where the dependencies among random variables are characterized by a dependency graph. We first define the notion of dependency graphs, which is a widely used model in probability, statistics, and combinatorics, see Erdős and Lovász (1975), Janson et al. (1988), Chen (1978) and Baldi and Rinott (1989) for some classical results.

Given a graph $G = (V, E)$, we say that random variables $\{X_i\}_{i \in V}$ are *G-dependent* if for any disjoint $S, T \subset V$ such that S and T are non-adjacent in G (that is, no edge in E has one endpoint in S and the other in T), random variables $\{X_i\}_{i \in S}$ and $\{X_j\}_{j \in T}$ are independent. See Fig. 3 for an example. Formally, we define the dependency graphs in the following.

Definition 2.4 (*Dependency graphs*) An undirected graph G is called a dependency graph of a random vector $\mathbf{X} = (X_1, \ldots, X_n)$ if

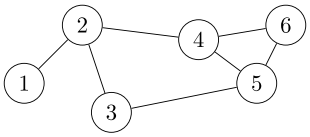


Fig. 3 A dependency graph G for random variables $\{X_i\}_{i \in [6]}$. Random variables $\{X_1, X_2\}$ and $\{X_5, X_6\}$ are independent, since disjoint vertex sets $\{1, 2\}$ and $\{5, 6\}$ are not adjacent in G

(b1) $V(G) = [n]$.
(b2) For all disjoint $I, J \subset [n]$, if I, J are not adjacent in G, then $\{X_i\}_{i\in I}$ and $\{X_j\}_{j\in J}$ are independent.

The above definition of dependency graphs is a strong version; there are ones with weaker assumptions, such as the one used in Lovász local lemma. Let K_n denote the complete graph on $[n]$, that is, every two vertices are adjacent. Then K_n is a dependency graph for any set of variables $\{X_i\}_{i\in[n]}$. Note that the dependency graph for a set of random variables may not be necessarily unique, and the sparser ones are the more interesting ones.

Here we introduce a widely-studied random process that generates dependent data whose dependency graph can be naturally constructed for illustration purposes. Consider a data-generating procedure modeled by the spatial Poisson point process, which is a Poisson point process on $\mathbb{R}^2$, see Linderman and Adams (2014) and Kirichenko and Van Zanten (2015) for discussions of using this process to model data collections in various machine learning applications. The number of points in each finite region follows a Poisson distribution, and the number of points in disjoint regions are independent. Given a finite set $\{U_i\}_{i=1}^n$ of regions in $\mathbb{R}^2$, let X_i be the number of points in region U_i for every $i \in [n]$. Then the graph $G\big([n], \{\{i,j\} : U_i \cap U_j \neq \emptyset\}\big)$ is a dependency graph of the random variables $\{X_i\}_{i=1}^n$.

An important property of the dependency graph, in view of the definition of fractional independent vertex covers, is that if we have a fractional independent vertex cover $\{(I_k, w_k)\}_{k\in[K]}$ of G, then we may decompose the sum of interdependent variables into a weighted sum of sums of independent variables.

Lemma 2.5 (Janson, 2004, Lemma 3.1) *Let G be a graph, and $\{(I_k, w_k)\}_{k\in[K]}$ be a fractional independent vertex cover of G. Let $\{u_i\}_{i\in V(G)}$ be a set of any numbers. Then*

$$\sum_{i\in V(G)} u_i = \sum_{i\in V(G)} \sum_{k=1}^{K} w_k \mathbf{1}_{\{i\in I_k\}} u_i = \sum_{k=1}^{K} w_k \sum_{i\in I_k} u_i, \tag{2.3}$$

where each $I_k \in \mathcal{I}(G)$ is an independent set. In particular, we have the following.

- By setting $u_i = 1$ for each $i \in V(G)$, we have

$$|V(G)|, = \sum_{k=1}^{K} w_k |I_k|. \tag{2.4}$$

- By letting $\{u_i\}_{i\in V(G)}$ be some G-dependent variables $\{X_i\}_{i\in V(G)}$, we have (2.3) becomes a weighted sum of independent random variables $\{X_i\}_{i\in I_k}$.

2.3 Concentration bounds for decomposable functions

Notice that McDiarmid's inequality applies to independent random variables. Janson (2004) derived a Hoeffding-like inequality for graph-dependent random variables by decomposing the sum into sums of independent variables. Janson's bound is a special case

of McDiarmid-type inequality tailored for interdependent random variables, especially when the function involves summation.

Theorem 2.6 (Janson's concentration inequality, 2004) *Let random vector* $\mathbf{X}$ *be G-dependent such that for every* $i \in V(G)$, *random variable* X_i *takes values in a real interval of length* $c_i \geqslant 0$. *Then, for every* $t > 0$,

$$\mathbb{P}\left(\sum_{i\in V(G)} X_i - \mathbb{E}\sum_{i\in V(G)} X_i \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\chi_f(G)\|\mathbf{c}\|_2^2}\right), \tag{2.5}$$

where $\mathbf{c} = (c_i)_{i\in V(G)}$ *and* $\chi_f(G)$ *is the fractional chromatic number of* G.

We will extend this result, and obtain similar concentration results under certain decomposability constraints for Lipschitz functions of graph-dependent random variables defined in Definition 2.7.

Definition 2.7 (*Decomposable* $\mathbf{c}$ *-Lipschitz functions*) Given a graph G on n vertices and a vector $\mathbf{c} = (c_i)_{i\in[n]} \in \mathbb{R}_+^n$, a function $f : \mathbf{\Omega} \to \mathbb{R}$ is *decomposable* $\mathbf{c}$*-Lipschitz* with respect to graph G if for all $\mathbf{x} = (x_1, \ldots, x_n) \in \mathbf{\Omega}$ and for all fractional independent vertex covers $\{(I_j, w_j)\}_j$ of G, there exist $(c_i)_{i\in I_j}$-Lipschitz functions $\{f_j : \mathbf{\Omega}_{I_j} \to \mathbb{R}\}_j$ such that

$$f(\mathbf{x}) = \sum_j w_j f_j(\mathbf{x}_{I_j}), \tag{2.6}$$

where for every set $V \subseteq [n]$, we write $\mathbf{\Omega}_V := \prod_{i\in V} \Omega_i$, and $\mathbf{x}_V := \{X_i\}_{i\in V}$.

Theorem 2.8 (Amini & Usunier, 2015; Usunier et al., 2005) *Let function* $f : \mathbf{\Omega} \to \mathbb{R}$ *be decomposable* $\mathbf{c}$*-Lipschitz, and* $\mathbf{\Omega}$*-valued random vector* $\mathbf{X}$ *be G-dependent. Then for* $t > 0$,

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\chi_f(G)\|\mathbf{c}\|_2^2}\right). \tag{2.7}$$

Remark 2.9 The chromatic number $\chi(G)$ of a graph G is the smallest number of colors needed to color the vertices of G such that no two adjacent vertices share the same color. Let $\Delta(G)$ denote the maximum degree of G. It is well-known that $\chi_f(G) \leqslant \chi(G) \leqslant \Delta(G) + 1$, (see, for example, Bollobás 1998). Thus in our bound (2.7), we can substitute $\chi_f(G)$ with $\chi(G)$ or $\Delta(G) + 1$, which may be easier to estimate in practice.

Proof of Theorem 2.8 Following the Cramér-Chernoff method (see, for example, Boucheron et al., 2013), we have for any $s > 0$ and $t > 0$,

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t\right) \leqslant e^{-st}\,\mathbb{E}\left[e^{s(f(\mathbf{X})-\mathbb{E}f(\mathbf{X}))}\right]. \tag{2.8}$$

Let $\{(I_j, w_j)\}_{j\in[J]}$ be a fractional independent vertex cover of the dependency graph G with

$$\sum_{j=1}^{J} w_j = \chi_f(G). \tag{2.9}$$

Utilizing the decomposition property of the Lipschitz function (2.6), the moment-generating function on the right-hand side of (2.8) can be written as

$$\mathbb{E}\left[e^{s(f(\mathbf{X})-\mathbb{E}f(\mathbf{X}))}\right] = \mathbb{E}\left[\exp\left(\sum_{j=1}^{J} sw_j(f_j(I_j)-\mathbb{E}f_j(I_j))\right)\right],$$

where each $f_j(I_j) = f_j(\mathbf{X}_{I_j})$ is some Lipschitz function of independent variables $\{X_i\}_{I\in I_j}$.

Now, let $\{p_1,\ldots,p_J\}$ be any set of J strictly positive reals that sum to 1. Since $\sum_{j=1}^{J}\omega_j/\chi_f(G)=1$ by (2.9), using the convexity of the exponential function and Jensen's inequality, we obtain that

$$\begin{aligned}\mathbb{E}\left[e^{s(f(\mathbf{X})-\mathbb{E}f(\mathbf{X}))}\right] &= \mathbb{E}\left[\exp\left(\sum_{j=1}^{J} p_j\frac{sw_j}{p_j}(f_j(I_j)-\mathbb{E}f_j(I_j))\right)\right]\\ &\leqslant \mathbb{E}\left[\sum_{j=1}^{J} p_j\exp\left(\frac{sw_j}{p_j}(f_j(I_j)-\mathbb{E}f_j(I_j))\right)\right]\\ &= \sum_{j=1}^{J} p_j\,\mathbb{E}\left[\exp\left(\frac{sw_j}{p_j}(f_j(I_j)-\mathbb{E}f_j(I_j))\right)\right],\end{aligned} \tag{2.10}$$

where the last step is by the linearity of expectation. Note that each subset I_j in summation (2.10) is an independent set, and therefore corresponds to independent variables. Hence applying Lemma 2.2 to each expectation that appears in the above summation gives

$$\sum_{j=1}^{J} p_j\,\mathbb{E}\left[\exp\left(\frac{sw_j}{p_j}(f_j(I_j)-\mathbb{E}f_j(I_j))\right)\right] \leqslant \sum_{j=1}^{J} p_j\exp\left(\frac{s^2w_j^2}{8p_j^2}\sum_{i\in I_j}c_i^2\right).$$

By rearranging terms in the exponential of the right-hand side of the inequality above and setting

$$p_j = \frac{w_j\sqrt{\sum_{i\in I_j}c_i^2}}{\sum_{j=1}^{J}\left(w_j\sqrt{\sum_{i\in I_j}c_i^2}\right)},$$

we have that

$$\begin{aligned}\sum_{j=1}^{J} p_j\exp\left(\frac{s^2w_j^2}{8p_j^2}\sum_{i\in I_j}c_i^2\right) &= \sum_{j=1}^{J} p_j\exp\left(\frac{s^2}{8}\left(\sum_{j=1}^{J} w_j\sqrt{\sum_{i\in I_j}c_i^2}\right)^2\right)\\ &= \exp\left(\frac{s^2}{8}\left(\sum_{j=1}^{J} w_j\sqrt{\sum_{i\in I_j}c_i^2}\right)^2\right),\end{aligned}$$

where the last equality is by recalling that the sum of p_i equals 1. By Cauchy–Schwarz inequality,

$$\left(\sum_{j=1}^{J} w_j \sqrt{\sum_{i\in I_j} c_i^2}\right)^2 = \left(\sum_{j=1}^{J} \sqrt{w_j}\sqrt{w_j \sum_{i\in I_j} c_i^2}\right)^2$$
$$\leqslant \left(\sum_{j=1}^{J} w_j\right)\left(\sum_{j=1}^{J} w_j \sum_{i\in I_j} c_i^2\right) = \chi_f(G) \sum_{i\in V(G)} c_i^2,$$

where the last equality is due to decomposition (2.3) and equation (2.9). The proof is then completed by choosing $s = 4t/(\chi_f(G)\sum_{i\in V(G)} c_i^2)$ in (2.8). □

2.4 Concentration bounds for general Lipschitz functions

We have demonstrated concentration results for functions with specific decomposable constraints. Moving forward, we extend our study to encompass more general Lipschitz functions. To begin with, we present concentration results for scenarios involving forest-dependence, wherein the dependency graphs are structured as forests. It is worth recalling that a forest is a disjoint union of trees.

Theorem 2.10 (Zhang et al., 2019; Zhang, 2022) *Let function $f : \mathbf{\Omega} \to \mathbb{R}$ be* **c**-*Lipschitz, and* $\mathbf{\Omega}$-*valued random vector* $\mathbf{X}$ *be G-dependent. If G is a disjoint union of trees* $\{T_i\}_{i\in[k]}$. *Then for* $t > 0$,

$$\mathbb{P}\left(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t\right) \leqslant \exp\left(-\frac{2t^2}{\sum_{i=1}^{k} c_{\min,i}^2 + \sum_{\{i,j\}\in E(G)}(c_i + c_j)^2}\right), \tag{2.11}$$

where $c_{\min,i} := \min\{c_j : j \in V(T_i)\}$ *for all* $i \in [k]$.

The proof of this theorem is by first properly ordering $\{X_i\}_{i\in V(G)}$ as $(X_i)_{i\in[n]}$, and rewriting $f(\mathbf{X}) - \mathbb{E}f(\mathbf{X})$ as a summation $\sum_{i\in[n]} V_i$, where

$$V_i := \mathbb{E}[f(\mathbf{X})|X_1, \ldots X_i] - \mathbb{E}[f(\mathbf{X})|X_1, \ldots X_{i-1}].$$

In the proof, each tree T_i is rooted by choosing the vertex with the minimum Lipschitz coefficient $\min\{c_j : j \in V(T_i)\}$ in that tree as the root. It can be shown that for some suitable ordering, each V_i ranges in an interval of length at most $c_i + c_j$, where j is the parent of i in the tree, or simply c_i (if i corresponds to a root vertex). The theorem then follows by applying the Chernoff-Cramér technique to $\sum_{i=1}^{n} V_i$. The detailed proof is a bit involved and can be found in Zhang (2022).

Remark 2.11 If random variables $(X_1, \ldots, X_n)$ are independent, then the empty graph $\overline{K}_n = ([n], \emptyset)$ is a valid dependency graphs for $\{X_i\}_{i\in[n]}$. In this case, inequality (2.11) gets reduced to the McDiarmid's inequality (2.2), since each vertex is treated as a tree.

If all Lipschitz coefficients are of the same value c, then the denominator of the exponent in (2.11) becomes $kc^2 + 4(n-k)c^2 = (4n-3k)c^2$, since the number of edges in the forest is $n-k$. The denominator in Janson's bound (2.5) is $2nc^2$, since the fractional chromatic number of any tree is 2. Thus if $k \geqslant 2n/3$, then bound (2.11) is tighter than Janson's concentration inequality (2.5).

2.4.1 Concentration for general graphs

In this subsection, we consider the concentration of general Lipschitz functions of variables whose dependency graph may not be a forest. This is by utilizing tree-partitions of the dependency graphs via vertex identifications, and then applying the forest-dependent results obtained.

Theorem 2.12 *Let function $f : \mathbf{\Omega} \to \mathbb{R}$ be $\mathbf{c}$-Lipschitz, and $\mathbf{\Omega}$-valued random vector $\mathbf{X}$ be G-dependent. Then for any $t > 0$,*

$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t) \leqslant \exp\left(-\frac{2t^2}{D(G,\mathbf{c})}\right),$$

where

$$D(G,\mathbf{c}) := \min_{(F,\{F_x\}_{x\in V(F)})\in \mathsf{TP}(G)} \left(\sum_{T\in\mathcal{T}(F)} \widetilde{c}^2_{\min,T} + \sum_{\{u,v\}\in E(F)} (\widetilde{c}_u + \widetilde{c}_v)^2 \right),$$

with $\widetilde{c}_u := \sum_{i\in F_u} c_i$ for all $u \in V(F)$ and $\widetilde{c}_{\min,T} := \min\{\widetilde{c}_i : i \in V(T)\}$ for all $T \in \mathcal{T}(F)$.

Proof For every $u \in V(F)$, we define a random vector $\mathbf{Y}_u = \{X_i\}_{i\in F_u}$, and treat each $\mathbf{Y}_u$ as a random variable. We then define a new random vector $\mathbf{Y} = (\mathbf{Y}_u)_{u\in V(F)}$, and let $g(\mathbf{Y}) = f(\mathbf{X})$. It is easy to check that g is $\widetilde{\mathbf{c}}$-Lipschitz by the triangle inequality, where $\widetilde{\mathbf{c}} = (\widetilde{c}_u)_{u\in V(F)}$. Hence the theorem immediately follows from Theorem 2.10. □

It is useful to define the notion of *forest complexity*, which depends only on the graph, especially when the Lipschitz coefficients are of the same order.

Definition 2.13 (*Forest complexity*) The forest complexity of a graph G is defined by

$$\Lambda(G) := \min_{(F,\{F_x\}_{x\in V(F)})\in \mathsf{TP}(G)} \left(\sum_{T\in\mathcal{T}(F)} \min_{u\in T} |F_u|^2 + \sum_{\{u,v\}\in E(F)} |F_u \cup F_v|^2 \right),$$

where the minimization is over all tree-partitions of G.

Remark 2.14 The width of a tree-partition is the maximum number of vertices in a bag. The tree-partition-width $\mathsf{tpw}(G)$ of G is the minimum width of a tree-partition of G. Let $F \in \mathsf{TP}(G)$ be the tree-partition with tree-partition width $\mathsf{tpw}(G)$. Then

$$\Lambda(G) \leqslant |\mathcal{T}(F)|\mathsf{tpw}(G)^2 + 4|E(F)|\mathsf{tpw}(G)^2 = (|V(F)| + 3|E(F)|)\mathsf{tpw}(G)^2,$$

since the number of disjoint trees in a forest F equals $|V(F)| - |E(F)|$. Upper bounds on tree-partition-width $\Lambda(G)$ can be obtained using treewidth and the maximum degree of G, and are beyond the scope of this paper, see Wood (2009) for more details.

If all the Lipschitz coefficients are of the same value, then Theorem 2.12 gets simplified.

Corollary 2.15 *Let function $f : \mathbf{\Omega} \to \mathbb{R}$ be Lipschitz with the same coefficient c, and $\mathbf{\Omega}$-valued random vector $\mathbf{X}$ be G-dependent. Then for $t > 0$,*

Fig. 4 A tree-partition of C_6

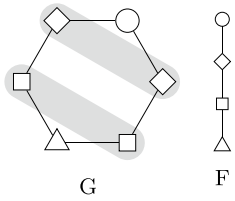
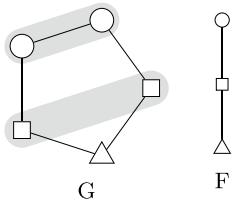


Fig. 5 A tree-partition of C_5



$$\mathbb{P}(f(\mathbf{X}) - \mathbb{E}f(\mathbf{X}) \geqslant t) \leqslant \exp\left(-\frac{2t^2}{\Lambda(G)c^2}\right). \tag{2.12}$$

Similar to the theorems derived above, Corollary 2.15 also gives an exponentially decaying bound on the probability of deviation. The rate of decay is determined by the Lipschitz coefficients of the function, and the forest complexity of the dependency graph. Intuitively, the closer the dependency graph is to a forest, the faster the deviation probability decays. This uncovers how the dependencies among random variables influence concentration.

2.4.2 Examples

Here we present several explicit examples to demonstrate and estimate the forest complexity where random variables are structured as graphs. All these examples naturally emerge in the context of random processes that are intricately intertwined within graph structures.

Example 2.16 (*G is a tree*) In this case, $\Lambda(G) \leqslant |E(G)|(1+1)^2 + 1 = 4n - 3$. We get an upper bound of $\Lambda(G)$ that is linear in the number of variables, which is comparable to Janson's concentration inequality up to some constant factor (see (2.5) with $\chi_f(G) = 2$ and Remark 2.11).

Example 2.17 (*G is a cycle C_n*) If n is even, a tree-partition is illustrated in Fig. 4, where the resulting forest is a path F of length $n/2$ with each gray belt representing a 'bag'. We will keep this convention for the rest of this paper.

By the illustrated tree-partition, $\Lambda(G) \leqslant 2 \times (1+2)^2 + (n/2 - 2)(2+2)^2 + 1 = O(n)$. When n is odd, according to the tree-partition shown in Fig. 5, $\Lambda(G) \leqslant (1+2)^2 + (\frac{n-1}{2} - 1)(2+2)^2 + 1 = O(n)$. Since $\chi_f \geqslant 2$ for cycles, our bound is again comparable to Janson's concentration inequality (2.5) up to some constant multiplicative factor.

Example 2.18 (*G is a grid*) Suppose G is a two-dimensional $(m \times m)$-grid. Then $n = m^2$. Considering the tree-partition illustrated in Fig. 6, we have

$$\Lambda(G) \leqslant 1 + 2\sum_{i=1}^{m}(2m-1)^2 = \frac{2}{3}m(2m+1)(2m-1) + 1 = O(m^3) = O(n^{\frac{3}{2}}).$$

3 Generalization for learning from graph-dependent data

We now apply the concentration bounds obtained above to derive generalization bounds for supervised learning from graph-dependent data. Let

$$\mathbf{S} := ((x_1, y_1), \ldots, (x_n, y_n)) \in (\mathcal{X} \times \mathcal{Y})^n$$

be a G-dependent training sample of size n, where $\mathcal{X}$ denotes the input space and $\mathcal{Y}$ denotes the set of labels. Let $\mathcal{D}$ be the underlying distribution of data on $\mathcal{X} \times \mathcal{Y}$. Note that the sample $\mathbf{S}$ contains dependent data with the same marginal distribution $\mathcal{D}$.

Further we fix some $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$ as a non-negative loss function. For any hypothesis $f : \mathcal{X} \to \mathcal{Y}$, the empirical error on sample $\mathbf{S}$ is defined by

$$\widehat{R}_{\mathbf{S}}(f) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i)).$$

For learning from dependent data, the generalization error can be defined in various ways. We adopt the following widely-used one (Hang & Steinwart, 2014; Lozano et al., 2006; Meir, 2000; Steinwart & Christmann, 2009)

$$R(f) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}[\ell(y, f(x))], \tag{3.1}$$

which assumes that the test data is independent of the training sample.

3.1 Generalization bounds via fractional Rademacher complexity

Our first approach is based on Rademacher complexity (Bartlett & Mendelson, 2002). This approach can be extended to accommodate interdependent data by utilizing the decomposition into independent sets described in Sect. 2.1.1.

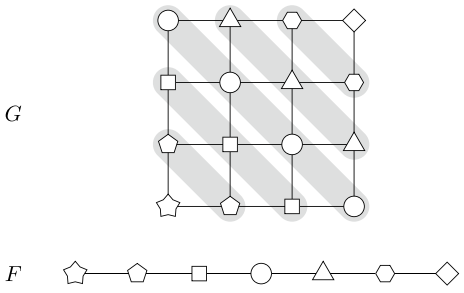


Fig. 6 A tree-partition of 4×4 gird

Definition 3.1 (*Fractional Rademacher complexity*, Usunier et al., 2005)
Let $\{(I_j, w_j)\}_j$ be a fractional independent vertex cover of a dependency graph G constructed over a training set $\mathbf{S}$ of size n, with $\sum_j w_j = \chi_f(G)$. Let $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be the hypothesis class. Then, the *empirical fractional Rademacher complexity* of $\mathcal{F}$ given $\mathbf{S}$ is defined by

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sum_j w_j \sup_{f \in \mathcal{F}} \left(\sum_{i \in I_j} \sigma_i f(x_i) \right) \right], \tag{3.2}$$

where $\boldsymbol{\sigma} = (\sigma_i)_{1 \leqslant i \leqslant n}$ denote a vector of n independent Rademacher variables, that is, $\mathbb{P}(\sigma_i = -1) = \mathbb{P}(\sigma_i = +1) = 1/2$ for each $i \in [n]$. Moreover, the *fractional Rademacher complexity* of $\mathcal{F}$ is defined by

$$\mathfrak{R}^{\star}(\mathcal{F}) = \mathop{\mathbb{E}}_{\mathbf{S}} \left[\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) \right].$$

Remark 3.2 In the i.i.d. situation, the set of singleton vertices is a valid fractional independent vertex cover, and the fractional Rademacher complexity (3.2) simplifies to the original empirical Rademacher complexity (Bartlett & Mendelson, 2002) defined by

$$\widehat{\mathfrak{R}}_{\mathbf{S}}(\mathcal{F}) = \frac{1}{n} \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{f \in \mathcal{F}} \left(\sum_{i \in [n]} \sigma_i f(x_i) \right) \right]. \tag{3.3}$$

Additionally, because the former is a sum of empirical Rademacher complexities, it enables one to get estimates by extending the properties of the empirical Rademacher complexity.

In the following, we give an example of a function class of linear functions with bounded-norm weight vectors, for which the empirical Rademacher averages can be bounded directly.

Theorem 3.3 *Let* $\mathcal{F} = \{x \mapsto \langle \mathbf{w}, \phi(x) \rangle : \|\mathbf{w}\| \leqslant B\}$ *be a class of linear functions with bounded weights in a feature space such that* $\|\phi(x)\| \leqslant \Gamma$ *for all* x. *Then*

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) \leqslant B\Gamma \sqrt{\frac{\chi_f(G)}{n}}. \tag{3.4}$$

Proof In view of the definition of the empirical fractional Rademacher complexity (3.2), by the linearity of expectation, we have

$$\begin{aligned} \widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) &= \frac{1}{n} \sum_j w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left[\sup_{\|\mathbf{w}\| \leqslant B} \left(\sum_{i \in I_j} \langle \mathbf{w}, \sigma_i \phi(x_i) \rangle \right) \right] \\ &\leqslant \frac{B}{n} \sum_j w_j \mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left\| \sum_{i \in I_j} \sigma_i \phi(x_i) \right\| \leqslant \frac{B}{n} \sum_j w_j \left(\mathop{\mathbb{E}}_{\boldsymbol{\sigma}} \left\| \sum_{i \in I_j} \sigma_i \phi(x_i) \right\|^2 \right)^{1/2}, \end{aligned}$$

where the first inequality is by noting $\|\mathbf{w}\| \leqslant B$, and applying Cauchy–Schwarz inequality to the inner product, and the second inequality is by Jensen's inequality.

As the Rademacher variables are independent, we have $\mathbb{E}[\sigma_i\sigma_k] = \mathbb{E}[\sigma_i]\,\mathbb{E}[\sigma_k] = 0$ for any distinct i, k. Hence we have

$$\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\left\|\sum_{i\in I_j}\sigma_i\phi(x_i)\right\|^2\right] = \mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sum_{i,k\in I_j}\sigma_i\sigma_k\langle\phi(x_i),\phi(x_k)\rangle\right] = \sum_{i\in I_j}\|\phi(x_i)\|^2,$$

and therefore,

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) \leqslant \frac{B}{n}\sum_j w_j\left(\sum_{i\in I_j}\|\phi(x_i)\|^2\right)^{1/2}.$$

Since we have $\|\phi(x_i)\| \leqslant \Gamma$ in the feature space, then

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) \leqslant \frac{B\Gamma}{n}\sum_j w_j\sqrt{|I_j|} = \frac{B\Gamma\chi_f(G)}{n}\sum_j\frac{w_j}{\chi_f(G)}\sqrt{|I_j|}.$$

By noticing that $\sum_j w_j/\chi_f(G) = 1$, using Jensen's inequality for the square root function yields

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\mathcal{F}) \leqslant \frac{B\Gamma\chi_f(G)}{n}\sqrt{\sum_j\frac{w_j}{\chi_f(G)}|I_j|} = \frac{B\Gamma\sqrt{\chi_f(G)}}{n}\sqrt{\sum_j w_j|I_j|}.$$

The result follows then by noting $\sum_j w_j|I_j| = n$ by (2.4). □

Remark 3.4 Note that ϕ could be the feature mapping corresponding to the last hidden layer of a neural network, or a kernel function. In particular, under the assumption of Theorem 3.3, let ϕ be a feature mapping associated to a kernel K such that $K(x,x) \leqslant \Gamma^2$ for all x. Then the standard Rademacher complexity of kernel-based hypotheses (Mohri et al., 2018, Theorem 6.12) gives that $\widehat{\mathfrak{R}}_{\mathbf{S}}(\mathcal{F}) \leqslant B\Gamma/\sqrt{n}$, and in comparison, our bound (3.4) has an additional factor $\sqrt{\chi_f(G)}$, which becomes exactly 1 as in Remark 3.2.

It is also worth noting that the fractional Rademacher complexity is defined for a given fractional cover. In general, our analysis holds for any optimal fractional cover; nevertheless, various cover selections may result in different bound values. Nonetheless, in practice, this influence is unlikely to have a significant impact.

We now obtain generalization bounds using the fractional Rademacher complexity.

Theorem 3.5 (Amini & Usunier, 2015; Usunier et al., 2005) *Given a sample* $\mathbf{S}$ *of size n with dependency graph G and a loss function* $\ell : \mathcal{Y}\times\widehat{\mathcal{Y}} \to [0,M]$. *Let* $\mathcal{F}$ *denote the hypothesis class. Then, for any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *we have, for all* $f \in \mathcal{F}$, *that*

$$R(f) \leqslant \widehat{R}_{\mathbf{S}}(f) + 2\mathfrak{R}^{\star}(\ell\circ\mathcal{F}) + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)}, \tag{3.5}$$

and

$$R(f) \leqslant \widehat{R}_{\mathbf{S}}(f) + 2\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\ell \circ \mathcal{F}) + 3M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{2}{\delta}\right)}, \tag{3.6}$$

where $\ell \circ \mathcal{F} = \{(x, y) \mapsto \ell(y, f(x)) \mid f \in \mathcal{F}\}$.

Proof For any $f \in \mathcal{F}$, we have $\widehat{R}_{\mathbf{S}}(f)$ is an unbiased estimator of $R(f)$, since the data points in the sample $\mathbf{S}$ are assumed to be G-dependent and have the same marginal distribution. Hence considering a G-dependent "ghost" sample $\mathbf{S}' = ((x'_1, y'_1), \ldots, (x'_n, y'_n))$ that is independently generated from the same distribution as $\mathbf{S}$, we have

$$\sup_{f\in\mathcal{F}}(R(f) - \widehat{R}_{\mathbf{S}}(f)) = \sup_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathbf{S}'}\widehat{R}_{\mathbf{S}'}(f) - \widehat{R}_{\mathbf{S}}(f)\right) = \sup_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathbf{S}'}\left[\widehat{R}_{\mathbf{S}'}(f) - \widehat{R}_{\mathbf{S}}(f)\right]\right).$$

Let $\{(I_j, w_j)\}_{j\in[J]}$ be a fractional independent vertex cover of the dependency graph G with $\sum_j w_j = \chi_f(G)$. By Jensen's inequality and the convexity of the supremum, we get

$$\begin{aligned}
\sup_{f\in\mathcal{F}}\left(\mathop{\mathbb{E}}_{\mathbf{S}'}\left[\widehat{R}_{\mathbf{S}'}(f) - \widehat{R}_{\mathbf{S}}(f)\right]\right) &\leqslant \mathop{\mathbb{E}}_{\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}(\widehat{R}_{\mathbf{S}'}(f) - \widehat{R}_{\mathbf{S}}(f))\right] \\
&= \mathop{\mathbb{E}}_{\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i\in[n]}(\ell(y'_i, f(x'_i)) - \ell(y_i, f(x_i)))\right)\right] \\
&= \frac{1}{n}\mathop{\mathbb{E}}_{\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}\left(\sum_{j=1}^{J} w_j \sum_{i\in I_j}(\ell(y'_i, f(x'_i)) - \ell(y_i, f(x_i)))\right)\right],
\end{aligned}$$

where the second equality is due to the decomposition (2.3).

Then by the sub-additivity of the supremum, we have

$$\sup_{f\in\mathcal{F}}(R(f) - \widehat{R}_{\mathbf{S}}(f)) \leqslant g(\mathbf{S}),$$

where $g(\mathbf{S})$ is defined by

$$g(\mathbf{S}) = \frac{1}{n}\mathop{\mathbb{E}}_{\mathbf{S}'}\left[\sum_{j=1}^{J} w_j \sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}(\ell(y'_i, f(x'_i)) - \ell(y_i, f(x_i)))\right)\right],$$

and satisfies $g(\mathbf{S}) = \sum_j w_j g_j(\mathbf{S})$, where for each j,

$$g_j(\mathbf{S}) := \frac{1}{n}\mathop{\mathbb{E}}_{\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}(\ell(y'_i, f(x'_i)) - \ell(y_i, f(x_i)))\right)\right].$$

Note that each function g_j has bounded difference M/n and satisfies (2.6), and therefore is a decomposable Lipschitz function. Then using Theorem 2.8, for all $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have

$$\sup_{f\in\mathcal{F}}(R(f)-\widehat{R}_{\mathbf{S}}(f)) \leqslant \mathop{\mathbb{E}}_{\mathbf{S}}[g(\mathbf{S})] + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)}$$
$$= \sum_{j=1}^{J}\frac{w_j}{n}\mathop{\mathbb{E}}_{\mathbf{S},\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}(\ell(y_i',f(x_i'))-\ell(y_i,f(x_i)))\right)\right] + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)}.$$

Note that

$$\mathop{\mathbb{E}}_{\mathbf{S},\mathbf{S}'}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}(\ell(y_i',f(x_i'))-\ell(y_i,f(x_i)))\right)\right]$$
$$= \mathop{\mathbb{E}}_{\mathbf{S},\mathbf{S}'}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i(\ell(y_i',f(x_i'))-\ell(y_i,f(x_i)))\right)\right],$$

since the introduction of Rademacher variables $\boldsymbol{\sigma}=(\sigma_i)_i$, uniformly taking values in $\{-1,+1\}$, does not change the expectation. Indeed, for $\sigma_i=+1$, the corresponding summand stays unaltered, and for $\sigma_i=-1$, the corresponding summand reverses sign, which is the same as flipping (x_i,y_i) and (x_i',y_i') between $\mathbf{S}$ and $\mathbf{S}'$. This change has no effect on the overall expectation as we are considering the expectation over $\mathbf{S}$ and $\mathbf{S}'$, and by noting that $\mathbf{S}$ and $\mathbf{S}'$ are independent and I_j is some independent set. Therefore, we have

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}(R(f)-\widehat{R}_{\mathbf{S}}(f))\\
&\leqslant \sum_{j=1}^{J}\frac{w_j}{n}\mathop{\mathbb{E}}_{\mathbf{S},\mathbf{S}'}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i(\ell(y_i',f(x_i'))-\ell(y_i,f(x_i)))\right)\right] + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)}\\
&\leqslant 2\sum_{j=1}^{J}\frac{w_j}{n}\mathop{\mathbb{E}}_{\boldsymbol{\sigma}}\mathop{\mathbb{E}}_{\mathbf{S}}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i\in I_j}\sigma_i(\ell(y_i,f(x_i)))\right)\right] + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)},
\end{aligned} \tag{3.7}$$

where the last step uses the sub-additivity of the supremum. Then in view of Definition 3.1 of $\widehat{\mathfrak{R}}_{\mathbf{S}}^{\star}$, we obtain

$$\sup_{f\in\mathcal{F}}(R(f)-\widehat{R}_{\mathbf{S}}(f)) \leqslant 2\mathop{\mathbb{E}}_{\mathbf{S}}\left[\widehat{\mathfrak{R}}_{\mathbf{S}}^{\star}(\ell\circ\mathcal{F})\right] + M\sqrt{\frac{\chi_f(G)}{2n}\log\left(\frac{1}{\delta}\right)}.$$

Therefore the first bound (3.5) follows from the definition of the supremum, that is, for all $f\in\mathcal{F}$,

$$R(f)-\widehat{R}_{\mathbf{S}}(f) \leqslant \sup_{f\in\mathcal{F}}\left(R(f)-\widehat{R}_{\mathbf{S}}(f)\right).$$

Note that

$$\widehat{\mathfrak{R}}^{\star}_{\mathbf{S}}(\ell \circ \mathcal{F}) = \sum_j w_j \left(\frac{1}{n} \mathop{\mathbb{E}}_{\sigma} \left[\sup_{f \in \mathcal{F}} \left(\sum_{i \in I_j} \sigma_i (\ell(y_i, f(x_i))) \right) \right] \right)$$

satisfies the condition of Theorem 2.8 with bounded difference M/n, and therefore concentrates around its expectation $\mathfrak{R}^{\star}(\ell \circ \mathcal{F})$. Then using the union bound with (3.5), yields the second bound (3.6). □

From Remark 3.2, Theorem 3.5 is a natural extension of the standard Rademacher generalization bounds when examples are identically and independently distributed (see, for example, Mohri et al., 2018, Theorem 3.3), as in this case, $\chi_f(G) = 1$.

Remark 3.6 To use the symmetrization technique in Eq. (3.7), the variables involved in the same summation need to be independent. Consequently, when extending the concept of Rademacher complexities to scenarios involving interdependent variables, it becomes necessary to decompose the set of random variables into independent sets. In this context, the fractional independent vertex cover $\{(I_j, w_j)\}_j$ with $\sum_k w_k = \chi_f(G)$ emerges as a pivotal tool for achieving an optimal decomposition, as $\chi_f(G)$ is the minimum of $\sum_k w_k$ over all fractional independent vertex covers.

3.2 Generalization bounds via algorithmic stability

This section establishes stability bounds for learning from graph-dependent data, using the concentration inequalities derived in the last section. Algorithmic stability has been used in the study of classification and regression to derive generalization bounds (Devroye & Wagner, 1979; Kearns & Ron, 1999; Kutin & Niyogi, 2002; Rogers & Wagner, 1978). A key advantage of stability bounds is that they are designed for specific learning algorithms, exploiting particular properties of the algorithms.

Since uniform stability was introduced in Bousquet and Elisseeff (2002), it has been among the most widely used notions of algorithmic stability. Given a training sample $\mathbf{S}$ of size n, for every $i \in [n]$, removing the i-th element from $\mathbf{S}$ results in a sample of size $n-1$, which is denoted by

$$\mathbf{S}^{\setminus i} := ((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}) \ldots, (x_n, y_n)).$$

A learning algorithm $\mathcal{A}$ is a function that maps the training set $\mathbf{S}$ onto a function $f^{\mathcal{A}}_{\mathbf{S}} : \mathcal{X} \to \mathcal{Y}$.

Definition 3.7 (*Uniform stability*, Bousquet & Elisseeff, 2002) Given an integer $n > 0$, the learning algorithm $\mathcal{A}$ is β_n-uniformly stable with respect to the loss function ℓ, if for any $i \in [n]$, $\mathbf{S} \in (\mathcal{X} \times \mathcal{Y})^n$, and $(x, y) \in \mathcal{X} \times \mathcal{Y}$, it holds that

$$|\ell(y, f^{\mathcal{A}}_{\mathbf{S}}(x)) - \ell(y, f^{\mathcal{A}}_{\mathbf{S}^{\setminus i}}(x))| \leqslant \beta_n. \tag{3.8}$$

Intuitively, small perturbations of the training sample have little effect on the learning for a stable learning algorithm.

Now, we begin our analysis by considering the difference between the empirical error and the generalization error of a learning algorithm $f^{\mathcal{A}}_{\mathbf{S}}$ trained over a G-dependent sample $\mathbf{S}$, formally defined by

$$\Phi_{\mathcal{A}}(\mathbf{S}) := R(f_{\mathbf{S}}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}). \tag{3.9}$$

The mapping $\Phi_{\mathcal{A}} : (\mathcal{X}\times\mathcal{Y})^n \to \mathbb{R}$ will play a critical role in estimating $R(f_{\mathbf{S}}^{\mathcal{A}})$ via stability. We will first bound the probability of the deviation of $\Phi_{\mathcal{A}}(\mathbf{S})$ from its expectation (Lemma 3.8), and then obtain an upper bound of expected value of $\Phi_{\mathcal{A}}(\mathbf{S})$ (Lemma 3.10).

Lemma 3.8 *Given a G-dependent sample* $\mathbf{S}$ *of size* n*, and a* β_n*-uniformly stable learning algorithm* $\mathcal{A}$*. Suppose the loss function* ℓ *is bounded by* M*. Then for any* $t > 0$*,*

$$\mathbb{P}(\Phi_{\mathcal{A}}(\mathbf{S}) - \mathbb{E}[\Phi_{\mathcal{A}}(\mathbf{S})] \geqslant t) \leqslant \exp\left(-\frac{2n^2t^2}{\Lambda(G)(4n\beta_n + M)^2}\right).$$

We prove the following lemma, which states that the Lipschitz coefficients of $\Phi_{\mathcal{A}}(\cdot)$ are all bounded by $4\beta_n + M/n$. Then Lemma 3.8 follows from Lemma 3.9 and Theorem 2.15, since the Lipschitz coefficients are all of the same value.

Lemma 3.9 *Given a* β_n*-uniformly stable learning algorithm* $\mathcal{A}$*, for any* $\mathbf{S}, \mathbf{S}' \in (\mathcal{X}\times\mathcal{Y})^n$ *that differ only in one entry, we have*

$$|\Phi_{\mathcal{A}}(\mathbf{S}) - \Phi_{\mathcal{A}}(\mathbf{S}')| \leqslant 4\beta_n + \frac{M}{n}.$$

Proof In the literature Bousquet and Elisseeff (2002), Lemma 3.9 was proved for the i.i.d. case, actually, the proof remains valid in our dependent setting. Assume that $\mathbf{S}$ and $\mathbf{S}'$ differ only in i-th entry, and denote $\mathbf{S}'$ as

$$\mathbf{S}^i := ((x_1, y_1), \ldots, (x_{i-1}, y_{i-1}), (x_i', y_i'), (x_{i+1}, y_{i+1}) \ldots, (x_m, y_m)),$$

such that the marginal distribution of (x_i', y_i') is also $\mathcal{D}$.

Notice that we do not require the data to be i.i.d., as samples are dependent, and have the same marginal probability distribution $\mathcal{D}$. To begin with, we bound $R(f_{\mathbf{S}}^{\mathcal{A}}) - R(f_{\mathbf{S}^i}^{\mathcal{A}})$ using the triangle inequality,

$$\begin{aligned}
\left|R(f_{\mathbf{S}}^{\mathcal{A}}) - R(f_{\mathbf{S}^i}^{\mathcal{A}})\right| &\leqslant \left|R(f_{\mathbf{S}}^{\mathcal{A}}) - R(f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}})\right| + \left|R(f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}) - R(f_{\mathbf{S}^i}^{\mathcal{A}})\right| \\
&= \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}}^{\mathcal{A}}(x))] - \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x))]\right| \\
&\quad + \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x))] - \mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}^i}^{\mathcal{A}}(x))]\right| \\
&= \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}}^{\mathcal{A}}(x)) - \ell(y, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x))]\right| \\
&\quad + \left|\mathop{\mathbb{E}}_{\mathcal{D}}[\ell(y, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x)) - \ell(y, f_{\mathbf{S}^i}^{\mathcal{A}}(x))]\right| \leqslant 2\beta_n,
\end{aligned}$$

where the last inequality is by the uniform stability defined by (3.8).

Then we bound $\widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}^i}(f_{\mathbf{S}^i}^{\mathcal{A}})$,

$$
\begin{aligned}
n\left|\widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}^i}(f_{\mathbf{S}^i}^{\mathcal{A}})\right| &= \left|\sum_{(x_j,y_j)\in\mathbf{S}} \ell(y_j, f_{\mathbf{S}}^{\mathcal{A}}(x_j)) - \sum_{(x_j,y_j)\in\mathbf{S}^i} \ell(y_j, f_{\mathbf{S}^i}^{\mathcal{A}}(x_j))\right| \\
&\leqslant \left|\ell(y_i, f_{\mathbf{S}}^{\mathcal{A}}(x_i)) - \ell(y_i', f_{\mathbf{S}^i}^{\mathcal{A}}(x_i'))\right| + \sum_{j\neq i}\left|\ell(y_j, f_{\mathbf{S}}^{\mathcal{A}}(x_j)) - \ell(y_j, f_{\mathbf{S}^i}^{\mathcal{A}}(x_j))\right| \\
&\leqslant \sum_{j\neq i}\left|\ell(y_j, f_{\mathbf{S}}^{\mathcal{A}}(x_j)) - \ell(y_j, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x_j))\right| + \sum_{j\neq i}\left|\ell(y_j, f_{\mathbf{S}^{\setminus i}}^{\mathcal{A}}(x_j)) - \ell(y_j, f_{\mathbf{S}^i}^{\mathcal{A}}(x_j))\right| \\
&\quad + \left|\ell(y_i, f_{\mathbf{S}}^{\mathcal{A}}(x_i)) - \ell(y_i', f_{\mathbf{S}^i}^{\mathcal{A}}(x_i'))\right| \leqslant 2n\beta_n + M,
\end{aligned}
$$

where the last inequality is by the uniform stability and the assumption that ℓ is bounded by M.

Combining the above bounds, by the triangle inequality, we have that

$$
\begin{aligned}
\left|\Phi_{\mathcal{A}}(\mathbf{S}) - \Phi_{\mathcal{A}}(\mathbf{S}^i)\right| &= \left|(R(f_{\mathbf{S}}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}})) - (R(f_{\mathbf{S}^i}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}^i}(f_{\mathbf{S}^i}^{\mathcal{A}}))\right| \\
&\leqslant \left|R(f_{\mathbf{S}}^{\mathcal{A}}) - R(f_{\mathbf{S}^i}^{\mathcal{A}})\right| + \left|\widehat{R}(f_{\mathbf{S}}^{\mathcal{A}}) - \widehat{R}_{\mathbf{S}^i}(f_{\mathbf{S}^i}^{\mathcal{A}})\right| \leqslant 4\beta_n + \frac{M}{n},
\end{aligned}
$$

which completes the proof. □

We are now in measure to bound the expectation of $\Phi_{\mathcal{A}}(\mathbf{S})$.

Lemma 3.10 *Let* $\mathbf{S}$ *be a* G*-dependent sample of size* n*. Suppose the maximum degree of* G *is* $\Delta = \Delta(G)$*. Let* $\mathcal{A}$ *be a* β_i*-uniformly stable learning algorithm for every* $i \in [n-\Delta, n]$*, and* $\beta_{n,\Delta} = \max_{i\in[0,\Delta]} \beta_{n-i}$*. Then we have*

$$
\mathbb{E}[\Phi_{\mathcal{A}}(\mathbf{S})] \leqslant 2\beta_{n,\Delta}(\Delta + 1).
$$

The proof of the lemma is based on iterative perturbations of the training sample $\mathbf{S}$, where a perturbation is essentially removing a data point from $\mathbf{S}$. The property of uniform stability of the algorithm guarantees that each perturbation causes a discrepancy up to $\beta_{n,\Delta}$, and in total $2(\Delta+1)$ perturbations have to be made to *eliminate* the dependency between a data point and the others.

We start with a technical lemma before the proof of Lemma 3.10.

Lemma 3.11 *Under the same assumptions in Lemma 3.10, we have*

$$
\max_{(x_i,y_i)\in\mathbf{S}} \mathop{\mathbb{E}}_{(x,y),\mathbf{S}}[\ell(y, f_{\mathbf{S}}^{\mathcal{A}}(x)) - \ell(y_i, f_{\mathbf{S}}^{\mathcal{A}}(x_i))] \leqslant 2\beta_{n,\Delta}(\Delta + 1).
$$

Proof For every $i \in [n]$, let $N_G(i)$ be the set of vertices adjacent to i in graph G, and suppose $N_G^+(i) = N_G(i) \cup \{i\} = \{j_1, \ldots, j_{n_i}\}$ with $j_{k-1} > j_k$. Define $\mathbf{S}^{(i,0)} = \mathbf{S}$ and for every $k \in [n_i]$, let $\mathbf{S}^{(i,k)}$ be obtained from $\mathbf{S}^{(i,k-1)}$ by removing the j_k-th entry. By the uniform stability of $\mathcal{A}$, for any $(x, y) \in \mathcal{X}\times\mathcal{Y}$ and $k \in [n_i]$, we have

$$
\left|\ell(y, f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x)) - \ell(y, f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x))\right| \leqslant \beta_{n,\Delta}.
$$

By a decomposition using a telescoping summation,

$$\ell(y,f_{\mathbf{S}}^{\mathcal{A}}(x)) = \sum_{k=1}^{n_i}(\ell(y,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x)) - \ell(y,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x)) + \ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)).$$

Similarly, we also get

$$\ell(y_i,f_{\mathbf{S}}^{\mathcal{A}}(x_i)) = \sum_{k=1}^{n_i}(\ell(y_i,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x_i)) - \ell(y_i,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x_i)) + \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i)).$$

Now we are ready to bound the difference

$$\begin{aligned}
&\ell(y,f_{\mathbf{S}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}}^{\mathcal{A}}(x_i)) \\
&= \sum_{k=1}^{n_i}\left((\ell(y,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x)) - \ell(y,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x))) - (\ell(y_i,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x_i)) - \ell(y_i,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x_i)))\right) \\
&\quad + \ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i)) \\
&\leqslant \sum_{k=1}^{n_i}|\ell(y,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x)) - \ell(y,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x))| \\
&\quad + \sum_{k=1}^{n_i}|\ell(y_i,f_{\mathbf{S}^{(i,k)}}^{\mathcal{A}}(x_i)) - \ell(y_i,f_{\mathbf{S}^{(i,k-1)}}^{\mathcal{A}}(x_i))| + \ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i)) \\
&\leqslant 2n_i\beta_{n,\Delta} + \ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i)).
\end{aligned}$$

Therefore, by noting that $n_i = |N_G^+(i)| \leqslant \Delta + 1$ for all i, we have

$$\begin{aligned}
&\mathop{\mathbb{E}}_{\mathbf{S},(x,y)}[\ell(y,f_{\mathbf{S}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}}^{\mathcal{A}}(x_i))] \\
&\leqslant \mathop{\mathbb{E}}_{\mathbf{S},(x,y)}[\ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i))] + 2n_i\beta_{n,\Delta} \\
&\leqslant \mathop{\mathbb{E}}_{\mathbf{S},(x,y)}[\ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i))] + 2\beta_{n,\Delta}(\Delta+1) \\
&= \mathop{\mathbb{E}}_{\mathbf{S},(x,y)}[\ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x))] - \mathop{\mathbb{E}}_{\mathbf{S}}[\ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i))] + 2\beta_{n,\Delta}(\Delta+1) \\
&= \mathop{\mathbb{E}}_{\mathbf{S}^{(i,n_i)},(x,y)}[\ell(y,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x))] - \mathop{\mathbb{E}}_{\mathbf{S}^{(i,n_i)},(x_i,y_i)}[\ell(y_i,f_{\mathbf{S}^{(i,n_i)}}^{\mathcal{A}}(x_i))] + 2\beta_{n,\Delta}(\Delta+1) \\
&= 2\beta_{n,\Delta}(\Delta+1),
\end{aligned}$$

where the last equality is because (x_i, y_i) and (x, y) are independent of $\mathbf{S}^{(i,n_i)}$, and have the same distribution. □

Now we are ready to prove Lemma 3.10.

Proof of Lemma 3.10 From the definition of $\Phi_{\mathcal{A}}(\mathbf{S})$ in (3.9), we have

$$\begin{aligned}
\mathop{\mathbb{E}}_{\mathbf{S}}[\Phi_{\mathcal{A}}(\mathbf{S})] &= \mathop{\mathbb{E}}_{\mathbf{S}}\left[\mathop{\mathbb{E}}_{(x,y)}[\ell(y,f_{\mathbf{S}}^{\mathcal{A}}(x))] - \frac{1}{n}\sum_{i=1}^{n}\ell(y_i,f_{\mathbf{S}}^{\mathcal{A}}(x_i))\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathop{\mathbb{E}}_{\mathbf{S},(x,y)}[\ell(y,f_{\mathbf{S}}^{\mathcal{A}}(x)) - \ell(y_i,f_{\mathbf{S}}^{\mathcal{A}}(x_i))] \leqslant 2\beta_{n,\Delta}(\Delta+1),
\end{aligned}$$

where the last inequality is by Lemma 3.11. □

Combining Lemmas 3.8 and 3.10 gives the following theorem, which upper-bounds the generalization error of learning algorithms trained over G-dependent training sets of size n.

Theorem 3.12 *Let* $\mathbf{S}$ *be a sample of size* n *with dependency graph* G. *Suppose the maximum degree of* G *is* Δ. *Assume that the learning algorithm* $\mathcal{A}$ *is* β_i*-uniformly stable for all* $i \in [n-\Delta, n]$. *Suppose the loss function* ℓ *is bounded by* M. *Let* $\beta_{n,\Delta} = \max_{i\in[0,\Delta]} \beta_{n-i}$. *For any* $\delta \in (0,1)$, *with probability at least* $1-\delta$, *it holds that*

$$R(f_{\mathbf{S}}^{\mathcal{A}}) \leqslant \widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}) + 2\beta_{n,\Delta}(\Delta+1) + \frac{4n\beta_n + M}{n}\sqrt{\frac{\Lambda(G)}{2}\log\left(\frac{1}{\delta}\right)}.$$

Remark 3.13 It is well known that for many learning algorithms, $\beta_n = O(1/n)$ (see, for example, Bousquet and Elisseeff 2002), in this case, we have that $\beta_{n,\Delta}(\Delta+1) \leqslant \beta_{n-\Delta}(\Delta+1) = O(\frac{\Delta}{n-\Delta})$, which vanishes asymptotically if $\Delta = o(n)$. The term $O\left(\sqrt{\Lambda(G)}/n\right)$ also vanishes asymptotically if $\Lambda(G) = o(n^2)$. We also observe that if the training data are i.i.d., Theorem 3.12 degenerates to the standard stability bound obtained in Bousquet and Elisseeff (2002), by setting $\Delta = 0$, $\beta_{n,\Delta} = \beta_n$, and $\Lambda(G) = n$.

4 Applications

In this section, we present three practical applications related to learning with interdependent data, for which we use the methodology presented in the previous sections to derive generalization bounds.

4.1 Bipartite ranking

The goal of bipartite ranking is to assign higher scores to instances of the positive class than the ones of the negative class (Agarwal & Niyogi, 2009; Freund et al., 2003). This framework corresponds to many applications of information retrieval such as recommender systems (Sidana et al., 2021), and uplift-modeling (Betlei et al., 2021), etc.

It has attracted a lot of interest in recent years since the empirical ranking error of a scoring function $h : \mathcal{X} \to \mathbb{R}$ over a training set $T := \left(x_i, y_i\right)_{1\leqslant i\leqslant m}$ with $y_i \in \{-1, +1\}$ defined by

$$\widehat{\mathcal{L}}_T(h) = \frac{1}{m_- m_+} \sum_{i:y_i=1} \sum_{j:y_j=-1} \mathbf{1}_{\{h(x_i)\leqslant h(x_j)\}}, \tag{4.1}$$

is equal to one minus the Area Under the ROC Curve (AUC) of h (see, for example, Cortes and Mohri 2004), where $m_- := \sum_{i=1}^m \mathbf{1}_{\{y_i=-1\}}$ and $m_+ := \sum_{i=1}^m \mathbf{1}_{\{y_i=1\}}$ are the number of negative and positive instances in the training set T respectively.

For two instances of different classes $(x, y), (x', y')$ in T such that $y \neq y'$, by considering the (unordered) pairs of examples $\{(x, y), (x', y')\}$, and the classifier of pairs f associated to a scoring function h defined by

$$f(x, x') = h(x) - h(x'),$$

we can rewrite the bipartite ranking loss (4.1) of h over T as the classification error of the associated f over the pairs of instances of different classes,

$$\widehat{R}_{\mathbf{S}}(f) = \widehat{\mathcal{L}}_T(h) = \frac{1}{n} \sum_{\{(x,y),(x',y')\}\in\mathbf{S}} \mathbf{1}_{\{z_{y,y'}f(x,x')\leqslant 0\}}, \tag{4.2}$$

where $n = m_- m_+$,

$$\mathbf{S} := \left\{(x,y),(x',y') : (x,y)\in T, (x',y')\in T, y\neq y'\right\}$$

is the set of n unordered pairs of examples from different classes in T, and

$$z_{y,y'} := 2\mathbf{1}_{\{y-y'>0\}} - 1.$$

Note that $z_{1,-1} = 1$ and $z_{-1,1} = -1$.

Let

$$T^+ := \{(x_i^+, 1) : i\in[m_+]\} \quad \text{and} \quad T^- := \{(x_j^-, -1) : j\in[m_-]\}$$

be the sets of positive and negative instances of T respectively. Then $T = T^+\cup T^-$. Without loss of generality, we assume that $m_+ \leqslant m_-$, which corresponds to the usual situation in information retrieval, where there are fewer positive (relevant) instances than negative (irrelevant) ones.

In this case, the independent covers of the corresponding dependency graph of $\mathbf{S}$ is $\{(I_k, 1)\}_{k\in\{1,\ldots,m_-\}}$, where

$$I_k = \left\{\left(x_i^+, x^-_{\sigma_{k,m_-}(i)}\right) : i\in[m_+]\right\},$$

with σ_{k,m_-} denoting the permutation that is defined by

$$\sigma_{k,m_-}(i) = \begin{cases} (k+i-1)(\mathrm{mod}\, m_-), & \text{if } (k+i-1)(\mathrm{mod}\, m_-)\neq 0 \\ m_-, & \text{otherwise.}\end{cases}$$

Figure 7 illustrates the dependency graph of a bipartite ranking problem with $m_+ = 2$ positive examples and $m_- = 3$ negative instances as well as its corresponding independent covers represented by dotted ellipsoids.

Remark 4.1 In the bipartite ranking, the dependent pairs of instances correspond to the edges of a complete bipartite graph K_{m_+,m_-}, since pairs are chosen with one positive instance and one negative instance, see Fig. 7 for illustration.

Given a graph G, the line graph of G has the edges of G as its vertices, with two vertices adjacent if the corresponding edges have a vertex in common in G. Then the dependency graph for pairs $\mathbf{S}$ is the line graph of K_{m_+,m_-}, known as an $m_+\times m_-$ Rook's graph, which is a Cartesian product of two complete graphs.

For bipartite ranking, it is easy to check that

$$\frac{\chi_f}{n} = \frac{\max(m_-, m_+)}{m_- m_+} = \frac{1}{\min(m_-, m_+)}.$$

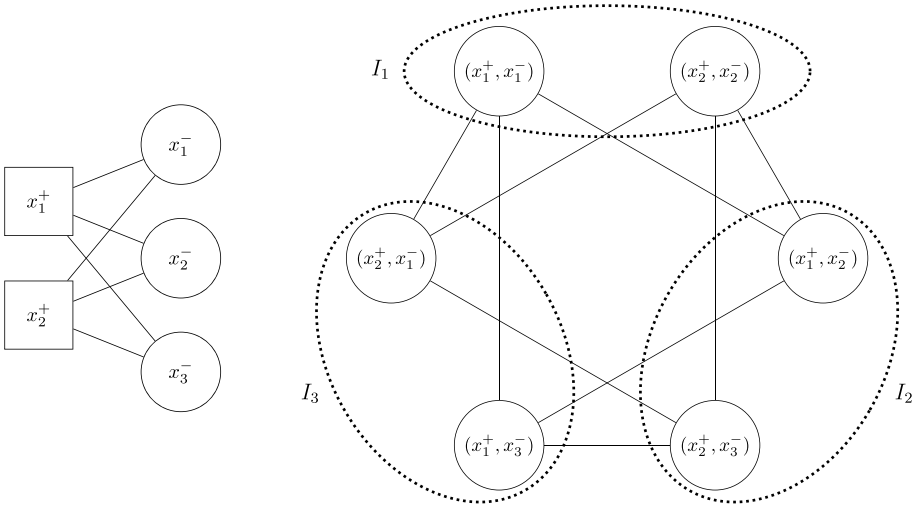


Fig. 7 The graph on the right is a dependency graph corresponding to a bipartite ranking problem with $m_+ = 2$ positive examples $T^+ = \{(x_1^+, 1), (x_2^+, 1)\}$; and $m_- = 3$ negative ones, $T^- = \{(x_1^-, -1), (x_2^-, -1), (x_3^-, -1)\}$. Each pair of examples from different classes corresponds to an edge of the complete bipartite graph $K_{2,3}$ on the left, and is represented by a vertex of the dependency graph on the right. Two pairs are adjacent in the dependency graph if they have an example in common. Fractional independent covers $\{(I_k, 1)\}_{1 \leqslant k \leqslant 3}$ are shown by dotted ellipsoids

Therefore by Theorems 3.3, 3.5, and Ledoux and Talagrand's contraction lemma (Ledoux & Talagrand, 1991, p.78 Corollary 3.17) that can be extended to fractional Rademacher complexities giving $\widehat{\mathfrak{R}}^\star_{\mathbf{S}}(\ell \circ \mathcal{F}) = 2\widehat{\mathfrak{R}}^\star_S(\mathcal{F})$, we can bound the generalization error of bipartite ranking as follows.

Corollary 4.2 *Let T be a training set composed of m_+ positive instances and m_- negative ones; and $\mathbf{S}$ the set of unordered pairs of examples from different classes in T. Then for any scoring function from $\mathcal{F} = \{f : (x, x') \mapsto \langle \mathbf{w}, \phi(x) - \phi(x') \rangle : \|\mathbf{w}\| \leqslant B\}$, where ϕ is a feature mapping with bounded norm such that $\|\phi(x) - \phi(x')\| \leqslant \Gamma$ for all (x, x'), and for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$R(f) \leqslant \widehat{R}_{\mathbf{S}}(f) + \frac{4B\Gamma}{\sqrt{m}} + 3\sqrt{\frac{1}{2m} \log\left(\frac{2}{\delta}\right)},$$

where $m = \min(m_-, m_+)$.

4.2 Multi-class classification

We now address the problem of mono-label multi-class classification, where the output space is a discrete set of labels $\mathcal{Y} = [K]$ with K classes. For the sake of

presentation, we denote an element of $\mathcal{X}\times\mathcal{Y}$ as $x^y := (x, y)$. For a class of predictor functions $\mathcal{H} = \{h : \mathcal{X}\times\mathcal{Y} \to \mathbb{R}\}$, let ℓ be the instantaneous loss of $h \in \mathcal{H}$ on example x^y defined by

$$\ell(y, h(x^y)) = \frac{1}{K-1} \sum_{y' \in \mathcal{Y}\setminus\{y\}} \mathbf{1}_{\{h(x^y) \leqslant h(x^{y'})\}}.$$

For any sample x, this loss function is the average number of classes, for which h assigns a higher score to the pairs constituted by x and any other classes that are not the true class of x. For a training set $T = \left(x_i^{y_i}\right)_{1\leqslant i\leqslant m}$ of size m, the corresponding empirical error of a function $h \in \mathcal{H}$ is

$$\widehat{\mathcal{L}}_T(h) = \frac{1}{m(K-1)} \sum_{i=1}^{m} \sum_{y' \in \mathcal{Y}\setminus\{y_i\}} \mathbf{1}_{\left\{h(x_i^{y_i}) \leqslant h(x_i^{y'})\right\}}. \tag{4.3}$$

Many multi-class classification algorithms like Adaboost.MR (Schapire & Singer, 1999) or the multiclass SVM (Weston & Watkins, 1998) aim to minimize a convex surrogate function of this loss.

Similar to the bipartite ranking case, by considering pairs $(x^y, x^{y'})$ with $y' \in \mathcal{Y}\setminus\{y\}$, constituted by the pairs x^y of an example and its class, and the pairs $x^{y'}$ of the same examples with all other classes, the classifier of pairs f associated to a function $h \in \mathcal{H}$ is defined by

$$f(x^y, x^{y'}) = h(x^y) - h(x^{y'}).$$

Then the empirical loss of a function h over T, can be written as the classification error of the associated f,

$$\widehat{R}_{\mathbf{S}}(f) = \widehat{\mathcal{L}}_T(h) = \frac{1}{n} \sum_{(x^y, x^{y'}) \in \mathbf{S}} \mathbf{1}_{\{z_{y,y'} f(x^y, x^{y'}) \leqslant 0\}}, \tag{4.4}$$

where $\mathbf{S} = \{(x^y, x^{y'}) : x^y \in T, x^{y'} \in T, y \neq y'\}$ is of size $n = m(K-1)$, and $z_{y,y'} = 2\mathbf{1}_{\{y>y'\}} - 1$. In this case, an independent cover of the corresponding dependency graph of $\mathbf{S}$ could be $\{(I_k, 1)\}_{k\in\{1,\ldots,K-1\}}$, where

$$I_k = \left\{\left(x_i^1, x_i^{k+1}\right) : i \in [m]\right\},$$

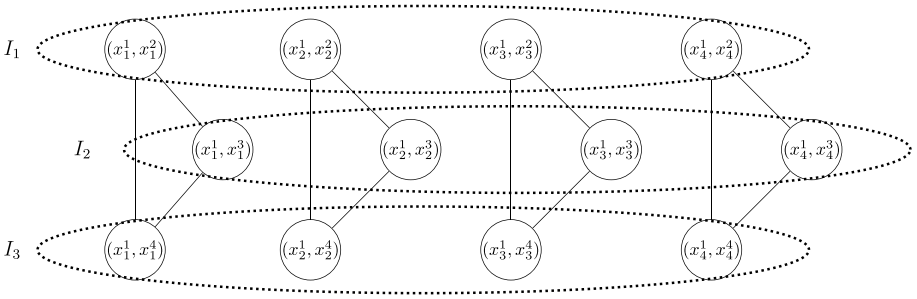


Fig. 8 The dependency graph for the multi-class classification problem with $m = 4$ examples and $K = 4$ classes is a vertex-disjoint union of 4 trianlges. Fractional independent covers $\{(I_k, 1)\}_{1\leqslant k\leqslant 3}$ are shown by dotted ellipsoids

with the corresponding fractional chromatic number $\chi_f = K - 1$.

Figure 8 illustrates a dependency graph for the multi-class classification problem with $m = 4$ and $K = 4$ as well as the corresponding fractional independent covers represented by dotted ellipsoids.

Similar to the bipartite ranking case, we have the following corollary based on the prior results.

Corollary 4.3 *Let T be a training set of K-label instances and of size m. Let $\mathbf{S}$ be the set of no-redundant pairs of examples from different classes in T. Then for any scoring functions from $\mathcal{F} = \{f : (x, x') \mapsto \langle \mathbf{w}, \phi(x) - \phi(x') \rangle : \|\mathbf{w}\| \leqslant B\}$, where ϕ is a feature mapping with the bounded norm such that $\|\phi(x) - \phi(x')\| \leqslant \Gamma$ for all (x, x'), and for any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$R(f) \leqslant \widehat{R}_{\mathbf{S}}(f) + \frac{4B\Gamma}{\sqrt{m}} + 3\sqrt{\frac{1}{2m} \log\left(\frac{2}{\delta}\right)}.$$

Remark 4.4 The loss function we considered (4.3) is normalized by $K - 1$, and we obtain a result that is comparable to the binary classification case. For a loss function based on margins, $\ell(y, h(x^y)) = h(x^y) - \max_{y' \neq y} h(x^{y'})$; the Rademacher complexity term grows in lockstep with the number of classes K.

4.3 Learning from *m*-dependent data

Here we consider learning from m-dependent data, and give a practical learning scenario. Suppose that there are linearly aligned locations, for example, real estate along a street. Let y_i be the observation at location i, for example, the house price. Let x_i denote the random variable modeling geographical effect at location i. Assume that x's are mutually independent and each y_i is geographically influenced by a neighborhood of size at most $2q + 1$. The goal is to learn to predict y from a sample $\{((x_{i-q}, \ldots, x_i, \ldots, x_{i+q}), y_i)\}_{i \in [n]}$, where n is the size of the sample. See Fig. 9 for an example.

This model accounts for the impact of local locations on house prices. Similar scenarios are frequently considered in spatial econometrics, and moving average processes in time series analysis, see Anselin (2013) for more examples.

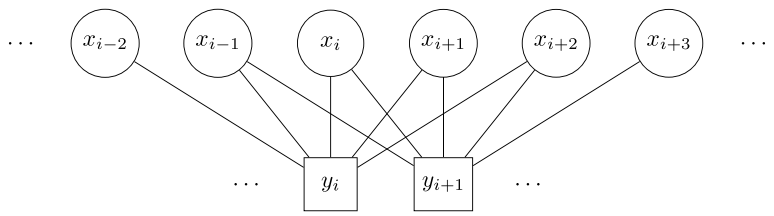


Fig. 9 Each observation y_i is geographically determined by a set of variables $\{x_j\}_{i-2 \leqslant j \leqslant i+2}$ of size 5. The sample $\{(\{x_j\}_{i-2 \leqslant j \leqslant i+2}, y_i)\}_i$ is 4-dependent

The above application is a special case of m-dependence. A sequence of random variables $\{X_i\}_{i=1}^n$ is said to be $f(n)$-dependent if subsets of variables separated by some distance greater than $f(n)$ are independent. This model was introduced by Hoeffding and Robbins (1948) and has been studied extensively (see, for example, Stein 1972; Chen 1975). This is usually the canonical application for the results based on the dependency graph model. A special case of $f(n)$-dependence when $f(n) = m$ is the following m-dependent model.

Definition 4.5 (*m-dependence*, Hoeffding and Robbins 1948) A sequence of random variables $\{X_i\}_{i=1}^n$ is m-dependent for some $m \geqslant 1$ if $\{X_j\}_{j=1}^i$ and $\{X_j\}_{j=i+m+1}^n$ are independent for all $i > 0$.

Figure 10 illustrates a dependency graph G for a 2-dependent sequence $\{X_i\}_i$, and its tree-partition. The illustration demonstrates the division of an m-dependent sequence into blocks of size m. Subsequently, these blocks are sequentially mapped to vertices of a path of length $\lceil n/m \rceil - 1$, as depicted in Fig. 10. This tree-partition shows that $\Lambda(G) \leqslant (\lceil n/m \rceil - 1)(m+m)^2 + m^2 \leqslant 4mn + m^2$.

Combining Theorem 3.12 and the above estimate of forest complexity gives the following.

Corollary 4.6 *Let* $\mathbf{S}$ *be an m-dependent sample of size n. Assume that the learning algorithm $\mathcal{A}$ is β_i-uniformly stable for any $i \in [n-2m, n]$. Suppose the loss function ℓ is bounded by M. For any $\delta \in (0,1)$, with probability at least $1-\delta$, it holds that*

$$R(f_{\mathbf{S}}^{\mathcal{A}}) \leqslant \widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}) + 2\beta_{n,2m}(2m+1) + (4n\beta_n + M)\sqrt{\frac{2m}{n}\left(1+\frac{m}{n}\right)\log\left(\frac{1}{\delta}\right)}.$$

Choose any uniformly stable learning algorithm in Bousquet and Elisseeff (2002) with $\beta_n = O(1/n)$, such as regularization algorithms in RKHS, etc., and apply to the above-mentioned house price prediction problem. Then for any fixed q, with high probability, Corollary 4.6 gives that $R(f_{\mathbf{S}}^{\mathcal{A}}) \leqslant \widehat{R}_{\mathbf{S}}(f_{\mathbf{S}}^{\mathcal{A}}) + O\left(\sqrt{\frac{1}{n}\log\left(\frac{1}{\delta}\right)}\right)$ for sufficiently large n, matching the stability bound in the i.i.d. case in Bousquet and Elisseeff (2002).

5 Concluding remarks

In this survey, we presented various McDiarmid-type concentration inequalities for functions of graph-dependent random variables. These concentration bounds were then used to obtain generalization error bounds for learning from graph-dependent samples via fractional Rademacher complexity and algorithm stability.

We also included some real practical applications of the methodology. Note that in our applications, the sample contains dependent data with the same marginal distribution, but this

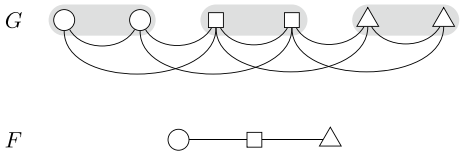


Fig. 10 A tree-partition of the dependency graph for 2-dependent variables

is not necessary and concentration inequalities derived are without this assumption, and therefore can be applied to situations where the distribution may change over time.

The dependency graphs used for our applications exhibit certain structural regularities and therefore we have explicit simple bounds. For applications under various other settings, we can still obtain meaningful bounds as long as we have suitable estimates of the fractional chromatic number or forest complexity. We will leave interested readers to investigate and find more applications.

There are various new directions that can be explored.

1. For dependent data, there are other definitions of the generalization error, such as the one specified in Kuznetsov and Mohri (2017) and Mohri and Rostamizadeh (2008, 2010). The connections between these and the one we used have been discussed in Mohri and Rostamizadeh (2008, 2010). It is a natural question whether our results can be adapted to this definition.
2. The dependency graph model we consider requires variables in disjoint non-adjacent subgraphs to be independent. There are some newly introduced dependency graph models such as weighted dependency graphs (Dousse & Féray, 2019; Féray, 2018), and the combination of mixing coefficients and dependency graphs (Lampert et al., 2018; Isaev et al., 2021). It would be interesting to use these new dependency graphs to obtain generalization bounds for learning under different dependent settings.
3. Recently, there are some new breakthroughs establishing sharper stability bounds (Bousquet et al., 2020; Feldman & Vondrak, 2019). It would be interesting to follow these results and to obtain sharper stability bounds for learning under graph-dependence.

Acknowledgements R.-R. Z. thanks David Wood for email communications on tree-partitions. The authors are sincerely grateful to the referees for carefully reading the manuscript and providing invaluable comments and suggestions, which led to a substantial improvement in the presentation.

Author contributions R.-R. Z.: the first and final draft, stability bound, and its applications. M.-R. A.: fractional Rademacher complexity bound, and its applications.

Funding The authors received no financial support for the research, authorship, and/or publication of this article.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The authors declare no conflict of interest.

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent for publication All authors participated in this study give the publisher the permission to publish this work.

References

Agarwal, S., & Niyogi, P. (2009). Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research, 10*(16), 441–474.

Amini, M. R., & Usunier, N. (2015). *Learning with partially labeled and interdependent data*. Springer.

Anselin, L. (2013). *Spatial econometrics: Methods and models* (Vol. 4). Springer.

Baldi, P., & Rinott, Y. (1989). On normal approximations of distributions in terms of dependency graphs. *The Annals of Probability, 17*(4), 1646–1650.

Bartlett, P. L., & Mendelson, S. (2002). Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research, 3*, 463–482.

Betlei, A., Diemert, E., & Amini, M. (2021). Uplift modeling with generalization guarantees. In *27th ACM SIGKDD conference on knowledge discovery and data mining* (pp. 55–65).

Bollobás, B. (1998). *Modern graph theory* (Vol. 184). Springer.

Boser, B., Guyon, I., & Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th annual workshop on computational learning theory (COLT'92)* (pp. 144–152).

Boucheron, S., Lugosi, G., & Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford University Press.

Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. *Journal of Machine Learning Research, 2*, 499–526.

Bousquet, O., Klochkov, Y., & Zhivotovskiy, N. (2020). Sharper bounds for uniformly stable algorithms. In *Conference on learning theory* (pp. 610–626). PMLR.

Chen, L. H. (1975). Poisson approximation for dependent trials. *The Annals of Probability,* 534–545

Chen, L. H. (1978). Two central limit problems for dependent random variables. *Probability Theory and Related Fields, 43*(3), 223–243.

Cortes, C., & Mohri, M. (2004). AUC optimization vs. error rate minimization. In *Advances in neural information processing systems*.

Dehling, H., & Philipp, W. (2002). Empirical process techniques for dependent data. In *Empirical process techniques for dependent data* (pp. 3–113). Springer.

Devroye, L., & Wagner, T. (1979). Distribution-free performance bounds for potential function rules. *IEEE Transactions on Information Theory, 25*(5), 601–604.

Dousse, J., & Féray, V. (2019). Weighted dependency graphs and the Ising model. *Annales de l'Institut Henri Poincaré D, 6*(4), 533–571.

Erdős, P., & Lovász, L. (1975). Problems and results on 3-chromatic hypergraphs and some related questions. *Infinite and Finite Sets, 10*(2), 609–627.

Feldman, V., & Vondrak, J. (2019). High probability generalization bounds for uniformly stable algorithms with nearly optimal rate. In *Conference on learning theory* (pp. 1270–1279). PMLR.

Féray, V. (2018). Weighted dependency graphs. *Electronic Journal of Probability, 23*.

Freund, Y., Iyer, R. D., Schapire, R. E., et al. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research, 4*, 933–969.

Halin, R. (1991). Tree-partitions of infinite graphs. *Discrete Mathematics, 97*(1–3), 203–217.

Hang, H., & Steinwart, I. (2014). Fast learning from α-mixing observations. *Journal of Multivariate Analysis, 127*, 184–199.

He, F., Zuo, L., & Chen, H. (2016). Stability analysis for ranking with stationary φ-mixing samples. *Neurocomputing, 171*, 1556–1562.

Hoeffding, W., & Robbins, H. (1948). The central limit theorem for dependent random variables. *Duke Mathematical Journal, 15*(3), 773–780.

Ibragimov, I. A. (1962). Some limit theorems for stationary processes. *Theory of Probability & its Applications, 7*(4), 349–382.

Isaev, M., Rodionov, I., & Zhang, R.R. et al (2021). Extremal independence in discrete random systems. arXiv preprint arXiv:2105.04917

Janson, S. (1988). Normal convergence by higher semiinvariants with applications to sums of dependent random variables and random graphs. *The Annals of Probability, 16*(1), 305–312.

Janson, S. (1990). Poisson approximation for large deviations. *Random Structures & Algorithms, 1*(2), 221–229.

Janson, S. (2004). Large deviations for sums of partly dependent random variables. *Random Structures & Algorithms, 24*(3), 234–248.

Janson, S., Łuczak, T., & Rucinski, A. (1988). An exponential bound for the probability of nonexistence of a specified subgraph in a random graph. Institute for Mathematics and its Applications (USA)

Kearns, M., & Ron, D. (1999). Algorithmic stability and sanity-check bounds for leave-one-out cross-validation. *Neural Computation, 11*(6), 1427–1453.

Kirichenko, A., & Van Zanten, H. (2015). Optimality of Poisson processes intensity learning with Gaussian processes. *The Journal of Machine Learning Research, 16*(1), 2909–2919.
Kontorovich, L. (2007). Measure concentration of strongly mixing processes with applications. Carnegie Mellon University.
Kontorovich, L. A., & Ramanan, K. (2008). Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability, 36*(6), 2126–2158.
Kutin, S., & Niyogi, P. (2002). Almost-everywhere algorithmic stability and generalization error. In *Proceedings of the eighteenth conference on uncertainty in artificial intelligence* (pp. 275–282). Morgan Kaufmann Publishers Inc.
Kuznetsov, V., & Mohri, M. (2017). Generalization bounds for non-stationary mixing processes. *Machine Learning, 106*(1), 93–117.
Lampert, C.H., Ralaivola, L., & Zimin, A. (2018). Dependency-dependent bounds for sums of dependent random variables. arXiv preprint arXiv:1811.01404
Ledoux, M., & Talagrand, M. (1991). *Probability in Banach spaces: Isoperimetry and processes*. Springer.
Linderman, S., & Adams, R. (2014). Discovering latent network structure in point process data. In *International conference on machine learning* (pp. 1413–1421).
Lozano, A. C., Kulkarni, S. R., & Schapire, R. E. (2006). Convergence and consistency of regularized boosting algorithms with stationary β-mixing observations. In: *Advances in neural information processing systems* (pp. 819–826).
McDiarmid, C. (1989). On the method of bounded differences. *Surveys in Combinatorics, 141*(1), 148–188.
Meir, R. (2000). Nonparametric time series prediction through adaptive model selection. *Machine Learning, 39*(1), 5–34.
Mohri, M., & Rostamizadeh, A. (2008). Stability bounds for non-i.i.d. processes. In *Advances in neural information processing systems* (pp. 1025–1032).
Mohri, M., & Rostamizadeh, A. (2009). Rademacher complexity bounds for non-i.i.d. processes. In *Advances in neural information processing systems* (pp. 1097–1104).
Mohri, M., & Rostamizadeh, A. (2010). Stability bounds for stationary φ-mixing and β-mixing processes. *Journal of Machine Learning Research, 11*, 789–814.
Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2018). *Foundations of machine learning*. MIT press.
Peña, V. H., & Giné, E. (1999). *Decoupling: From dependence to independence*. Springer.
Ralaivola, L., & Amini, M. R. (2015). Entropy-based concentration inequalities for dependent variables. In *International conference on machine learning* (pp. 2436–2444).
Ralaivola, L., Szafranski, M., & Stempfel, G. (2010). Chromatic PAC-Bayes bounds for non-iid data: Applications to ranking and stationary β-mixing processes. *Journal of Machine Learning Research, 11*, 1927–1956.
Rogers, W. H., & Wagner, T. J. (1978). A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics,* 506–514.
Rosenblatt, M. (1956). A central limit theorem and a strong mixing condition. *Proceedings of the National Academy of Sciences of the United States of America, 42*(1), 43.
Schapire, R. E., & Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine Learning, 37*(3), 297–336.
Seese, D. (1985). Tree-partite graphs and the complexity of algorithms. In *International conference on fundamentals of computation theory* (pp.412–421) Springer.
Sidana, S., Trofimov, M., Horodnytskyi, O., et al. (2021). User preference and embedding learning with implicit feedback for recommender systems. *Data Mining Knowledge Discovery, 35*(2), 568–592.
Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proceedings of the sixth Berkeley symposium on mathematical statistics and probability, volume 2: Probability theory*. The Regents of the University of California.
Steinwart, I., & Christmann, A. (2009). Fast learning from non-i.i.d. observations. In *Advances in neural information processing systems* (pp. 1768–1776).
Usunier, N., Amini, M. R., & Gallinari, P. (2005). Generalization error bounds for classifiers trained with interdependent data. *Advances in Neural Information Processing Systems, 18*, 1369–1376.
Volkonskii, V., & Rozanov, Y. A. (1959). Some limit theorems for random functions. I. *Theory of Probability & its Applications, 4*(2), 178–197.
Weston, J., & Watkins, C. (1998). *Multi-class support vector machines*.
Wood, D. R. (2009). On tree-partition-width. *European Journal of Combinatorics, 30*(5), 1245–1253.
Yu, B. (1994). Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability,* 94–116.
Zhang, R. R. (2022). When Janson meets McDiarmid: Bounded difference inequalities under graph-dependence. *Statistics & Probability Letters, 181*(109), 272.

Zhang, R. R., Liu, X., Wang, Y., & Wang, L. (2019). McDiarmid-type inequalities for graph-dependent variables and stability bounds. *Advances in Neural Information Processing Systems, 32*, 10890–10901.