On the ERM Principle with Networked Data

Yuanhong Wang¹, Yuyi Wang², Xingwu Liu³, Juhua Pu¹
¹ The State Key Laboratory of Software Development Environment, Beihang University, Beijing, China
² Disco Group, ETH Zurich, Switzerland
³ Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China

Abstract

Networked data, in which every training example involves two objects and may share some common objects with others, is used in many machine learning tasks such as learning to rank and link prediction. A challenge of learning from networked examples is that target values are not known for some pairs of objects. In this case, neither the classical i.i.d. assumption nor techniques based on complete U-statistics can be used. Most existing theoretical results of this problem only deal with the classical empirical risk minimization (ERM) principle that always weights every example equally, but this strategy leads to unsatisfactory bounds. We consider general weighted ERM and show new universal risk bounds for this problem. These new bounds naturally define an optimization problem which leads to appropriate weights for networked examples. Though this optimization problem is not convex in general, we devise a new fully polynomial-time approximation scheme (FPTAS) to solve it.

1 Introduction

“No man is an island, entire of itself ...”, the beginning of a well-known poem by the 17th century English poet John Donne, might be able to explain why social networking websites are so popular. These social media not only make communications convenient and enrich our lives but also bring us data, of an unimaginable amount, that is intrinsically networked. Social network data nowadays is widely used in research on social science, network dynamics, and as an inevitable fate, data mining and machine learning (Scott 2017). Similar examples of networked data such as traffic networks (Min and Wynter 2011), chemical interaction networks (Szklarczyk et al. 2014), citation networks (Dawson et al. 2014) also abound throughout the machine learning world.

Admittedly, many efforts have been made to design practical algorithms for learning from networked data, e.g., (Liben-Nowell and Kleinberg 2007, Mackassy and Provost 2007, Li et al. 2016, Garcia-Duran et al. 2016). However, not many theoretical guarantees of these methods have been established, which is the main concern of this paper. More specifically, this paper deals with risk bounds of classifiers trained with networked data (CLANET) whose goal is to train a classifier with examples in a data graph G. Every vertex of G is an object and described by a feature vector X ∈ X that is drawn independently and identically (i.i.d.) from an unknown distribution, while every edge corresponds to a training example whose input is a pair of feature vectors (X, X′) of the two ends of this edge and whose target value Y is in {0, 1}.

A widely used principle to select a proper model from a hypothesis set is Empirical Risk Minimization (ERM). Papa, Bellet, and Clémençon (2016) establish risk bounds for ERM on complete data graphs, and the bounds are independent of the distribution of the data. These bounds are of the order O(log(n)/n), where n is the number of vertices in the complete graph. However, in practice it is very likely that one cannot collect examples for all pairs of vertices and then G is usually incomplete, thus techniques based on complete U-processes in (Papa, Bellet, and Clémençon 2016) cannot be applied and the risk bounds of the order O(log(n)/n) are no longer valid in this setting. By generalizing the moment inequality for U-processes to the case of incomplete graphs, we prove novel risk bounds for the incomplete graph.

Usually, every training example is equally weighted (or unweighted) in ERM, which seems much less persuasive when the examples are networked, in particular when the graph is incomplete. But, most existing theoretical results of learning from networked examples are based on the unweighted ERM (Usunier, Amini, and Gallinari 2006, Ralaivola, Szafranski, and Stempfel 2009), and their bounds are of the order O(√χ^*(D_G)/m) where D_G is the line graph of G and χ^* is the fractional chromatic number of D_G (see Section A in the appendix) and m is the number of training examples. In order to improve this bound, Wang, Guo, and Ramon (2017) propose weighted ERM which adds weights to training examples according to the data graph, and show that the risk bound for weighted ERM can be of the order O(1/√ν^*(G)) where ν^*(G) is the fractional matching number of G, so using weighted ERM networked data can be more effectively exploited than the equal weighting method, as basic graph theory tells us ν^*_G ≥ m/χ^*(D_G). However, Wang, Guo, and Ramon (2017) (in fact, Usunier, Amini, and Gallinari (2006) and Ralaivola, Szafranski, and Stempfel (2009) also) assume that any two examples can be arbitrarily correlated if they share a vertex, which cannot lead to an O(log(n)/n) bound when the graph is complete.

We show that the “low-noise” condition, also called the Mammen-Tsybakov noise condition (Mammen and Tsybakov 1999)
bound on incomplete graphs for tasks such as pairwise ranking and graph reconstruction that satisfy the “low-noise” condition. Under this condition, the method proposed in (Wang, Guo, and Ramon 2017) does not work (see Section 6.1).

Before we show new risk bounds and new weighting methods, we present the following three aspects to convey some intuitions.

Line Graphs Compared to Janson’s decomposition which is based on line graphs, our method utilizes the additional dependency information in the data graph $G$. For example, the complete line graph with three vertices (i.e., triangle) corresponds to two different data graphs, as illuminated in Figure 1. Hence, line graph based methods ignore some important information in the data graph. This negligence makes it unable to improve bounds, no matter whether considering weighted ERM or not (see Section A.1 in the appendix). In Section 6.2, we show that our bounds are tighter than that of line graph based methods.

Asymptotic Risk As mentioned by Wang, Guo, and Ramon (2017), if several examples share a vertex, then we are likely to put less weight on them because the influence of this vertex to the empirical risk should be bounded. Otherwise, if we treat every example equally, then these dependent examples may dominate the training process and lead to the risk bounds that do not converge to 0 (see the example in Section 6.2).

Uniform Bounds Ralaivola and Amini (2015) prove an entropy-base concentration inequality for networked data using Janson’s decomposition, but the assumption there is usually too restrictive to be satisfied (see Section A.2 in appendix). To circumvent this problem, our method uses the “low-noise” condition (also used in (Papa, Bellet, and Clémencçon 2016)) to establish uniform bounds, in absence of any restrictive condition imposed on the data distribution.

3 Preliminaries

In this section, we begin with the detailed probabilistic framework for CLANET, and then give the definition of weighted ERM on networked examples.

3.1 Problem Statement

Consider a graph $G = (V,E)$ with a vertex set $V = \{1, \ldots, n\}$ and a set of edges $E \subseteq \{(i,j): 1 \leq i \neq j \leq n\}$. For each $i \in V$, a continuous random variable (r.v.) $X_i$, taking its values in a measurable space $\mathcal{X}$, describes features of vertex $i$. The $X_i$’s are i.i.d. r.v.’s following some unknown distribution $P_X$. Each pair of vertices $(i,j) \in E$ corresponds to a networked example whose input is a pair $(X_i, X_j)$ and target value is $Y_{i,j} \in \mathcal{Y}$. We focus on binary classification in this paper, i.e., $\mathcal{Y} = \{0, 1\}$. Moreover, the distribution of target values only depends on the features of the vertices it contains but does not depend on features of other vertices, that is, there is a probability distribution $P_{\mathcal{Y}|X^2}$ such that for
This paper considers an optimization problem, where we show that the "low-noise" constant bounds for the excess risk of the Bayes rule. For this problem, we select a classifier noted by \( \hat{f} \) by minimizing the empirical risk \( \ell(f, X, Y) \) defined as

\[
\ell(f, X, Y) = \frac{1}{m} \sum_{(i,j) \in E} \ell(r, X_{i,j}).
\]

As mentioned, the "low-noise" condition can lead to tighter risk bounds. For this problem, we show that the "low-noise" condition for the i.i.d. part of the Hoeffding decomposition (Hoeffding 1948) of its excess risk can be always obtained if the problem is symmetric (see Lemma 2).

\[\ell(r, X_{i,j}),\]
**Definition 4** (covering numbers). Let $(F, L_p)$ be a metric space with $L_p$-pseudometric. We define the covering number $N(F, L_p, \varepsilon)$ be the minimal $l \in \mathbb{N}$ such that there exist $l$ disks in $F$ with radius $\varepsilon$ covering $F$. If the context is clear, we simply denote $N(F, L_p, \varepsilon)$ by $N_p(F, \varepsilon)$.

In this paper, we focus on the $L_\infty$ covering number $N_\infty(F, \varepsilon)$ and suppose that it satisfies the following assumption.

**Assumption 1.** There exists a nonnegative number $\beta < 1$ and a constant $K$ such that $\log N_\infty(F, \varepsilon) \leq K \varepsilon^{-\beta}$ for all $\varepsilon \in (0, 1]$.

Similar to (Massart and Nédélec 2006) and (Rejchel 2012), we restrict to $\beta < 1$, whereas in the empirical process theory this exponent usually belongs to $[0, 2)$. This restriction is needed to prove Lemma 1, which involves the integral of $\log N_\infty(F, \varepsilon)$ through 0. Dudley (1974), Korostelev and Tsybakov (1993) and Mannen and Tsybakov (1995) presented various examples of classes $F$ satisfying Assumption 1. We also refer interested readers to (Mannen and Tsybakov 1998, p. 1813) for more concrete examples of hypothesis classes with smooth boundaries satisfying Assumption 1.

### 4.2 Risk Bounds

Now we are ready to show the tighter risk bounds for weighted empirical risk by the following theorem.

**Theorem 1** (risk bounds). Let $r_w$ be a minimizer of the weighted empirical risk $L_w$ over a class $R$ that satisfies Assumption 1. There exists a constant $C > 0$ such that for all $\delta \in (0, 1]$, with probability at least $1 - \delta$, the excess of $r_w$ satisfies

$$L(r_w) - L^* \leq 2\left(\inf_{r \in R} L(r) - L^*\right) + \frac{K' C \log(1/\delta)}{(1 - \beta)^{2/(\beta+1)}} \left(\|w\|_1^{\beta/(1+\beta)} + \max\left(\|w\|_2, \|w\|_\infty(\log(1/\delta))^{1/2}, \|w\|_\infty(\log(1/\delta))\right)\right)$$

where $\|w\|_\text{max} = \max_i \sqrt{\sum_{j,(i,j) \in E} w_{i,j}^2}$ and $K' = \max(K, \sqrt{K}, K^{1/(1+\beta)})$.

According to Theorem 1, if the parameter $\delta$ is greater than the value $\exp\left(-\min(\|w\|_2 / \|w\|_\infty, \|w\|_2^2 / \|w\|_\text{max}^2)\right)$, then the risk bounds above are of the order $O\left(1/(\|w\|_1^{1/(1+\beta)}) + \|w\|_2 / \|w\|_1 \right)$. In this case, our bounds are tighter than $O(1/\sqrt{\|w\|_1})$, which are fractional matching and $0 < \beta < 1$. If $G$ is complete and every example is equally weighted, the bounds of the order $O((1/n)^{1/(1+\beta)})$ achieve the same results as in (Papa, Bellet, and Clémenceon 2016).\(^1\)

**Remark.** Theorem 1 provides universal risk bounds no matter what the distribution of the data is. The factor of $2$ in front of the approximation error $\inf_{r \in R} L(r) - L^*$ has no special meaning and can be replaced by any constant larger than $1$ with a cost of increasing the constant $C$. Wang, Guo, and Ramon (2017) obtain risk bounds that has a factor $1$ in front of the approximation error part, but in their result the bound is $O(1/\sqrt{\|w\|_1})$. Hence, Theorem 1 improves their results if the approximation error does not dominate the other terms in the bounds.

In the rest of this section, we outline the main ideas to obtain this result. We first define $\tilde{g}_r(x_1, x_2, y_1, 2) := \ell(r, x_1, x_2, y_1, 2) - \ell(r^*, x_1, x_2, y_1, 2)$ for every $(x_1, x_2, y_1, 2) \in X \times X \times Y$ and let $\Lambda(r) := L(r) - L^* = \mathbb{E}[\tilde{g}_r(X_1, X_2, Y_1, 2)]$ be the excess risk with respect to the Bayes rule. Its empirical estimate by weighted ERIM is

$$\Lambda_w(r) = L_w(r) - L_w(r^*) = \frac{1}{\|w\|_1} \sum_{(i,j) \in E} w_{i,j} \tilde{g}_r(X_i, X_j, Y_{i,j}).$$

By Hoeffding’s decomposition (Hoeffding 1948), for all $r \in R$, one can write

$$\Lambda_w(r) = T_w(r) + U_w(r) + \tilde{U}_w(r),$$

where $T_w(r) = \Lambda(r) + \frac{2}{\|w\|_1} \sum_{i=1}^n \sum_{j,(i,j) \in E} w_{i,j} h_r(X_i)$ is a weighted average of i.i.d. random variables with $h_r(X_i) = \mathbb{E}[\tilde{g}_r(X_i, X_j, Y_{i,j}) | X_i] - \Lambda(r)$.

$$U_w(r) = \frac{1}{\|w\|_1} \sum_{(i,j) \in E} w_{i,j} \tilde{h}_r(X_i, X_j)$$

is a weighted degenerated (i.e., the symmetric kernel $h_r(x_1, x_2)$ such that $\mathbb{E}[\tilde{h}_r(X_i, X_j) | X_1 = x_1] = 0$ for all $x_1 \in X$) $U$-statistic $\tilde{h}_r(X_i, X_j) = \mathbb{E}[\tilde{g}_r(X_i, X_j, Y_{i,j}) | X_i, X_j] - \Lambda(r) - h_r(X_i) - h_r(X_j)$ and

$$\tilde{U}_w(r) = \frac{1}{\|w\|_1} \sum_{(i,j) \in E} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j}),$$

with a degenerated kernel $\tilde{h}_r(X_i, X_j, Y_{i,j}) = \tilde{g}_r(X_i, X_j, Y_{i,j}) = \mathbb{E}[\tilde{g}_r(X_i, X_j, Y_{i,j}) | X_i, X_j]$. In the following, we bound the three terms $T_w, U_w$, and $\tilde{U}_w$ in (5) respectively.

**Lemma 1** (uniform approximation). Under the same assumptions as in Theorem 1, for any $\delta \in (0, 1/\varepsilon)$, we have with probability at least $1 - \delta$,

$$\sup_{r \in R} |U_w(r)| \leq \max(K, \sqrt{K}) C_1 \frac{\max(\|w\|_2 \log(1/\delta), \|w\|_\infty(\log(1/\delta))^{1/2})}{1 - \beta} \frac{\max(\|w\|_2 \log(1/\delta), \|w\|_\infty(\log(1/\delta))^{1/2})}{\|w\|_1}$$

and

$$\sup_{r \in R} |\tilde{U}_w(r)| \leq \max(K, \sqrt{K}) C_2 \left(\frac{\|w\|_2}{1 - \beta} \frac{\max(\|w\|_2 \log(1/\delta), \|w\|_\infty(\log(1/\delta))^{1/2})}{\|w\|_1} + \max(\|w\|_\infty(\log(1/\delta))^{1/2}, \|w\|_\infty(\log(1/\delta))^{1/2})\right)$$

where $C_1, C_2 < +\infty$ are constants.

---

\(^1\)They consider the same range of $\delta$. 

To prove Lemma 1, we show that $U_w(r)$ and $\tilde{U}_w(r)$ can be bounded by Rademacher chaos using classical symmetrization and randomization tricks combined with the decoupling method. We handle these Rademacher chaos by generalizing the moment inequality for $U$-statistics in (Clémencó, Lugosi, and Vayatis 2008). Specifically, we utilize the moment inequalities from (Boucheron et al. 2005) to convert them into a sum of simpler processes, which can be bounded by the metric entropy inequality for Khinchine-type processes (see Arcones and Giné 1993, Proposition 2.6) and Assumption 1. The detailed proofs can be found in Section C in the appendix.

Lemma 1 shows that the contribution of the degenerated parts $U_w(r)$ and $\tilde{U}_w(r)$ to the excess risk can be bounded. This implies that minimizing $A_w(r)$ is approximately equivalent to minimizing $T_w(r)$ and thus $r_w$ is a $\rho$-minimizer of $T_w(r)$ in the sense that $T_w(r_w) \leq \rho + \inf_{r \in R} T_w(r)$. In order to analyze $T_w(r)$, which can be treated as a weighted empirical risk on i.i.d. examples, we generalize the results in (Massart and Nédélec 2006) (see Section B in the appendix). Based on this result, tight bounds for the excess risk with respect to $T_w(r)$ can be obtained if the variance of the excess risk is controlled by its expected value. By Lemma 2, $T_w(r)$ fulfills this condition, which leads to Lemma 3.

**Lemma 2** (condition leads to “low-noise”, (Papa, Bellet, and Clémencó 2016, Lemma 2)).

If the learning problem is symmetric, then

$$\var{E}[q_r(X_1, X_2, Y_1, Z_1) | X_1] \leq \Lambda(r)$$

holds for any distribution $P$ and any function $r \in R$.

**Lemma 3** (risk bounds for i.i.d. examples). Suppose that $r'$ is a $\rho$-minimizer of $T_w(r)$ in the sense that $T_w(r') \leq \rho + \inf_{r \in R} T_w(r)$ and $R$ satisfies Assumption 1, then there exists a constant $C$ such that for all $\delta \in (0, 1]$, with probability at least $1 - \delta$, the risk of $r'$ satisfies

$$\Lambda(r') \leq 2 \inf_{r \in R} \Lambda(r) + 2\rho + \frac{C K^{1/(1+\beta)} \log(1/\delta)}{||w||_1 (1 - \beta)^2 (1+\beta)}.$$

With Lemma 1 Lemma 3, now we are ready to prove Theorem 1.

**Proof of Theorem 1.** Let us consider the Hoeffding decomposition (5) of $A_w(r)$ that is minimized over $r \in R$. The idea of this proof is that the degenerate parts $U_w(r)$ and $\tilde{U}_w(r)$ can be bounded by Lemma 1. Therefore, $r_w$ is an approximate minimizer of $T_w(r)$, which can be handled by Lemma 3.

Let $A$ be the event that

$$\sup_{r \in R} |U_w(r)| \leq \kappa_1,$$

where

$$\kappa_1 = \frac{C_1}{1 - \beta} \max\left( \frac{\|w\|_2 \log(1/\delta)}{\|w\|_1}, \frac{\|w\|_{\max} \log(1/\delta)}{\|w\|_1}^{3/2}, \frac{\|w\|_{\infty} \log(1/\delta)}{\|w\|_1} \right)$$

for an appropriate constant $C_1$. Then by Lemma 1, $P[A] \geq 1 - \delta/4$. Similarly, let $B$ be the event that

$$\sup_{r \in R} |\tilde{U}_w(r)| \leq \kappa_2.$$

where

$$\kappa_2 = \frac{C_2}{1 - \beta} \left( \frac{\|w\|_2^2 + \max\left( \frac{\|w\|_{\max} \log(1/\delta)}{\|w\|_1}, \frac{\|w\|_{\infty} \log(1/\delta)}{\|w\|_1} \right)^3/2}{\|w\|_1} \right)$$

for an appropriate constant $C_2$. Then $P[B] \geq 1 - \delta/4$.

By (5), it is clear that, if both $A$ and $B$ happen, $r_w$ is a $\rho$-minimizer of $T_w(r)$ over $r \in R$ in the sense that the difference between the value of this latter quantity at its minimum and $r_w$ is at most $(\kappa_1 + \kappa_2)$. Then, from Lemma 3, with probability at least $1 - \delta/2$, $r_w$ is a $(\kappa_1 + \kappa_2)$-minimizer of $T_w(r)$, which the result follows.

An intuition obtained from our result is how to choose weights for networked data. By Theorem 1, to obtain tight risk bounds, we need to maximize $\|w\|_1$ (under the constraint that this weight vector is a fractional matching), which resembles the result of (Wang, Guo, and Ramon 2017) (but they only need to maximize $\|w\|_1$ and this is why they end in the $O(1/\sqrt{\gamma(G)}$ bound), while making $\|w\|_2, \|w\|_{\max}, \|w\|_{\infty}$ as small as possible, which appears to suggest putting nearly average weights on examples and vertices respectively. These two objectives, maximizing $\|w\|_1$ and minimizing $\|w\|_2, \|w\|_{\max}, \|w\|_{\infty}$, seem to contradict each other. In the next section, we discuss how to solve this problem.

## 5 Weighting Vector Optimization

In this section, we first formulate the optimization problem that minimizes the risk bounds in Theorem 1. Although this optimization problem is not convex unless $\beta = 0$, which usually means that there is no general efficient way to solve it, we devise a fully polynomial-time approximation scheme (FPTAS) to solve it.

**Definition 5** (FPTAS). An algorithm $A$ is a FPTAS for a minimization problem $\Pi$, if for any input $\mathcal{I}$ of $\Pi$ and $\epsilon > 0$, $A$ finds a solution $s$ in time polynomial in both the size of $\mathcal{I}$ and $1/\epsilon$ that satisfies

$$f_{\Pi}(s) \leq (1 + \epsilon) \cdot f_{\Pi}(s^*),$$

where $f_{\Pi}$ is the (positive) objective function of $\Pi$ and $s^*$ is an optimal solution for $\mathcal{I}$.

### 5.1 Optimization Problem

According to Theorem 1, given a graph $G$, $\beta \in (0, 1)$ and $\delta \in (0, 1]$, one can find a good weighting vector with tight
risk bounds by solving the following program:

$$\min_w \frac{1}{\|w\|_1} \left( \|w\|_1^{\beta/(1+\beta)} + \max \left( \|w\|_2, \|w\|_{\infty} (\log(1/\delta))^{1/2} \right) \right)$$

s.t. \( \forall (i,j) \in E, w_{i,j} \geq 0 \) and \( \forall i, \sum_{j:(i,j) \in E} w_{i,j} \leq 1 \)

(7)

To get rid of the fraction of norms in the program above, we consider a distribution \( p \) on edges \( p_{i,j} := w_{i,j}/\|w\|_1 \) and then \( \|w\|_1 \leq 1/\max_{i=1,\ldots,n} \sum_{j:(i,j) \in E} p_{i,j} \). Every distribution \( p \) corresponds to a valid weighting vector \( w \). By introducing two auxiliary variables \( a \) and \( b \), solving the original program (7) is equivalent to solving

$$\min_{a,b,p} a^{1/(1+\beta)} + b$$

s.t. \( \forall (i,j) \in E, p_{i,j} \geq 0 \)

\( \forall (i,j) \in E, p_{i,j} \log(1/\delta) - b \leq 0 \)

\( \forall i, \sum_{j:(i,j) \in E} p_{i,j} - a \leq 0 \)

\( \forall i, \left( \sum_{j:(i,j) \in E} p_{i,j}^2 \log(1/\delta) \right)^{1/2} - b \leq 0 \)

\( \|p\|_2 - b \leq 0 \) and \( \sum_{(i,j) \in E} p_{i,j} = 1 \)

(8)

Note that the constraints are all convex. If \( \beta = 0 \), e.g., the hypothesis set is finite, then the objective function becomes linear and thus (8) is a convex optimization problem that can be solved by some convex optimization method (see e.g., (Boyd and Vandenberghe 2004)) such as interior-point method.

If \( \beta > 0 \), the objective function is not convex any more. In fact, the program (8) becomes a concave problem that may be optimized globally by some complex algorithms (Benson 1995, Hoffman 1981) that often need tremendous computation. Instead, one may only need to approximate it using some efficient methods, e.g., Concave-Convex Procedure (Yuille 2001) and Coordinate Descent (Wright 2015). However, these methods lack in complexity analysis and may lead to a local optimum.

### 5.2 A Fully Polynomial-time Approximation Scheme

To solve the program (8) efficiently, we propose Algorithm 1 and show that it is a fully polynomial-time approximation scheme for (8).

**Theorem 2.** Algorithm 1 is a FPTAS for the program (8).

**Proof.** We first analyze the running time of this algorithm.

---

**Algorithm 1** FPTAS for weighting vector optimization.

**Input:** \( \epsilon, \beta, \delta \) and a graph \( G \) that contains \( n \) vertices and \( m \) edges.

**Output:** An approximate optimal weighting vector \( \mathbf{p} \) for the program (8).

1. Solve the following linear program (LP) efficiently\(^2\), and obtain an \( \epsilon \)-approximation \( a_{\min} \):

$$\min_{a,p} a \quad \text{s.t.} \quad \forall (i,j) \in E, p_{i,j} \geq 0$$

\( \forall i, \sum_{j:(i,j) \in E} p_{i,j} - a \leq 0 \)

$$\sum_{(i,j) \in E} p_{i,j} = 1 \quad (9)$$

2. Let \( \text{Grid} := \{a_{\min} + i \cdot \epsilon/(1+\beta) / n \mid i \in \mathbb{N} \text{ and } i \leq \frac{n(1-a_{\min})}{\epsilon(1+\beta)} \} \) and \( \text{Solutions} := \emptyset \).

3. for \( a \in \text{Grid} \) do

4. Use some efficient interior-point method to obtain an \( \epsilon \)-approximation of the following program and add the solution \((a, b, p)\) into \( \text{Solutions} \).

$$\min_{b,p} b \quad \text{s.t.} \quad \forall (i,j) \in E, p_{i,j} \geq 0$$

\( \forall i, \sum_{j:(i,j) \in E} p_{i,j} \log(1/\delta) - b \leq 0 \)

\( \forall i, \sum_{j:(i,j) \in E} p_{i,j} - a \leq 0 \)

$$\|p\|_2 - b \leq 0 \text{ and } \sum_{(i,j) \in E} p_{i,j} = 1 \quad (10)$$

5. return the vector \( \mathbf{p} \) which makes \( a^{1/(1+\beta)} + b \) smallest from \( \text{Solutions} \).

---

Note that \( 1/n \leq a \leq 1 \) if the graph is not empty. In Algorithm 1, we first divide the problem into at most

$$\frac{1 - 1/n}{\epsilon(1+\beta)/n} = \frac{n - 1}{\epsilon(1+\beta)}$$

convex programs, each of which produces an \( \epsilon \)-approximate solution by some interior-point method. Since interior-point method is FPTAS for convex problems (Boyd and Vandenberghe 2004), solving each of these programs needs polynomial time in the problem size \( m + n \) and \( 1/\epsilon \). Thus, the complexity of Algorithm 1 is also polynomial in \( m + n \) and \( 1/\epsilon \).

Now we show that this algorithm indeed results in an \( \epsilon \)-approximation of this optimal solution.

For any optimal solution \((a^* ,b^* ,\mathbf{p}^*)\), if \( a^* \) achieves minimum for the program (9), we can find \( a' \) in \( \text{Grid} \) (actually

\(^2\)For example, some interior-point method.
When graph $G$ is complete, weighting all examples equally gives the best risk bound, as all the terms $\max_{i=1,...,n} \sum_{j: (i,j) \in E} p_{i,j}$, $\|p\|_2$, $\|p\|_{\text{max}}$ and $\|p\|_\infty$ achieve minimum. Compared to the results in (Wang, Guo, and Ramon 2017), our theory puts additional constrains on $\|p\|_2$, $\|p\|_{\text{max}}$ and $\|p\|_\infty$ which encourages weighting examples fairly in this case, as illustrated in Figure 1. Besides, this scheme, which coincides with $U$-statistics that average the basic estimator applied to all sub-samples, produces the smallest variance among all unbiased estimators (Hoefding 1948).

6.2 Equal Weighting

Let us discuss further the equal weighting scheme that gives every example the same weight. Denote by $\Delta(G)$ the maximum degree of $G$ (note that this is not the maximum degree of $D_G$) and let $p_{i,j} = 1/m$ (recall that $m$ is the number of examples) for all $(i, j) \in E$. According to program (8), using equal weighting scheme, the risk bounds are of the order

$$O\left(\left(\frac{\Delta(G)}{m}\right)^{1/(1+\beta)} + \frac{1}{\sqrt{m}}\right),$$

if $\delta \in (\exp(-m/\Delta(G)), 1]$. In some cases, such as bounded degree graphs and complete graphs, this scheme provides reasonable risk bounds. Note that $\Delta(G)$ is smaller than the maximum size of cliques in its corresponding line graph $D_G$ and $\chi^*(D_G)$ is larger than the maximum size of cliques in $D_G$; these bounds above are always better than the bounds of the order $O\left(\sqrt{\chi^*(D_G)/m}\right)$ built by Janson’s decomposition.

However, as argued in Section 2, one can construct examples to illustrate that if we use the equal weighting strategy when $\Delta(G)$ is large (e.g., if it is linear to $m$), the risk bounds (14) are very large and do not converge to 0, while this problem can be solved by simply using a better weighting strategy.

Example 2. Consider a data graph with $|E| = m \gg 1$ and $E$ consists of $m/2$ disjoint edges and $m/2$ edges sharing a common vertex, then $\Delta(G) = m/2$. Using the equal weighting scheme, the risk bounds are of the order $O(1)$ that is meaningless. A much better weighting scheme of this case is to weight the examples of disjoint edges with $2/(m+2)$ while weight the examples of adjacent edges with $4/m(m+2)$, which provides risk bounds of the order $O\left((1/m)^{1/(1+\beta)} + \sqrt{1/m}\right)$.

7 Conclusion

In this paper, we consider weighted ERM of the symmetric CLANET problem and establish new universal risk bounds under the “low-noise” condition. These new bounds are tighter in the case of incomplete graphs and can be degenerate to the known tightest bound when graphs are complete. Based on this result, one can train a classifier with a better risk bound by putting proper weights on training examples. We propose an efficient algorithm to obtain the approximate optimal weighting vector and prove that the algorithm is a FPTAS for the weighting vector optimization problem. Finally, we discuss two cases to show the merits of our new risk bounds. In particular, we show that, in terms of these risk bounds, weighting all examples equally is optimal for complete graphs but not always good for incomplete graphs.
References

Arcones, M. A., and Gine, E. 1993. Limit theorems for u-processes. The Annals of Probability 1494–1542.
Benson, H. P. 1995. Concave Minimization: Theory, Applications and Algorithms. Springer US.
Biau, G., and Bleakley, K. 2006. Statistical inference on graphs. Statistics & Decisions 24(2):209–232.
Boucheron, S.; Bousquet, O.; Lugosi, G.; Massart, P.; et al. 2005. Moment inequalities for functions of independent random variables. The Annals of Probability 33(2):514–560.
Bousquet, O. 2002. A bennett concentration inequality and its application to suprema of empirical processes. Comptes Rendus Mathematique 334(6):495–500.
Boyd, S., and Vandenberghe, L. 2004. Convex optimization. Cambridge university press.
Clémençon, S.; Lugosi, G.; and Vayatis, N. 2008. Ranking and empirical minimization of u-statistics. Annals of Statistics 36(2):844–874.
Cucker, F., and Zhou, D. X. 2007. Learning theory: an approximation theory viewpoint, volume 24. Cambridge University Press.
Dawson, S.; Gašević, D.; Siemens, G.; and Joksimovic, S. 2014. Current state and future trends: A citation network analysis of the learning analytics field. In Proceedings of the Fourth International Conference on Learning Analytics and Knowledge, 231–240.
De la Peña, V., and Giné, E. 2012. Decoupling: from dependence to independence. Springer Science & Business Media.
Dudley, R. M. 1974. Metric entropy of some classes of sets with differentiable boundaries. Journal of Approximation Theory 10(3):227–236.
Garcia-Duran, A.; Bordes, A.; Usunier, N.; and Grandvalet, Y. 2016. Combining two and three-way embedding models for link prediction in knowledge bases. Journal of Artificial Intelligence Research 55:715–742.
Hoeffding, W. 1948. A class of statistics with asymptotically normal distribution. The Annals of Mathematical Statistics 293–325.
Hoffman, K. L. 1981. A method for globally minimizing concave functions over convex sets. Mathematical Programming 20(1):22–32.
Janson, S. 2004. Large deviations for sums of partly dependent random variables. Random Structures & Algorithms 24(3):234–248.
Korostelev, A. P., and Tsybakov, A. B. 1993. Minimax Theory of Image Reconstruction. Springer-Verlag.
Li, J.; Hu, X.; Wu, L.; and Liu, H. 2016. Robust unsupervised feature selection on networked data. In Proceedings of the 2016 SIAM International Conference on Data Mining, 387–395.
Liben-Nowell, D., and Kleinberg, J. 2007. The link-prediction problem for social networks. Journal of the Association for Information Science and Technology 58(7):1019–1031.
Liu, T.-Y. 2009. Learning to rank for information retrieval. Foundations and Trends® in Information Retrieval 3(3):225–331.
Macskassy, S. A., and Provost, F. 2007. Classification in networked data: A toolkit and a univariate case study. Journal of Machine Learning Research 8(May):935–983.
Mammen, E., and Tsybakov, A. B. 1995. Asymptotical minimax recovery of sets with smooth boundaries. Annals of Statistics 23(2):502–524.
Mammen, E., and Tsybakov, A. B. 1998. Smooth discrimination analysis. Annals of Statistics 27(6):1808–1829.
Massart, P., and Nédélec, É. 2006. Risk bounds for statistical learning. Annals of Statistics 2326–2366.
Min, W., and Wynter, L. 2011. Real-time road traffic prediction with spatio-temporal correlations. Transportation Research Part C: Emerging Technologies 19(4):606–616.
Papa, G.; Bellet, A.; and Clémençon, S. 2016. On graph reconstruction via empirical risk minimization: Fast learning rates and scalability. In Advances in Neural Information Processing Systems, 694–702.
Ralaivola, L., and Amini, M.-R. 2015. Entropy-based concentration inequalities for dependent variables. In International Conference on Machine Learning, 2436–2444.
Ralaivola, L.; Szafranski, M.; and Stempfel, G. 2009. Chromatic pac-bayes bounds for non-iid data. In Artificial Intelligence and Statistics, 416–423.
Rejchel, W. 2012. On ranking and generalization bounds. Journal of Machine Learning Research 13(May):1373–1392.
Scott, J. 2017. Social network analysis. Sage.
Szklarczyk, D.; Franceschini, A.; Wyder, S.; Forsslund, K.; Heller, D.; Huerta-Cepas, J.; Simonovic, M.; Roth, A.; Santos, A.; Tsafou, K. P.; et al. 2014. String v10: protein–protein interaction networks, integrated over the tree of life. Nucleic Acids Research 43(D1):D447–D452.
Tsybakov, A. B. 2004. Optimal aggregation of classifiers in statistical learning. Annals of Statistics 135–166.
Usunier, N.; Amini, M.-R.; and Gallinari, P. 2006. Generalization error bounds for classifiers trained with interdependent data. In Advances in neural information processing systems, 1369–1376.
Vapnik, V., and Chervonenkis, A. Y. 1971. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability & Its Applications 16(2):264–280.
Wang, Y.; Guo, Z.-C.; and Ramon, J. 2017. Learning from networked examples. In Proceedings of the 28th International Conference on Algorithmic Learning Theory, to appear.
Wright, S. J. 2015. Coordinate descent algorithms. Mathematical Programming 151(1):3–34.
Yuille, A. L. 2001. The concave-convex procedure (cccp). In Advances in Neural Information Processing Systems 14, 915 – 936.
Appendix

This appendix is organized as follows. Section A provides new risk bounds for weighted ERM following the Janson’s decomposition and show that it cannot improve prior results. Section B establishes universal risk bounds for weighted ERM on i.i.d. examples under the "low-noise" condition and proves upper bounds involved in covering number for weighted empirical processes. Section C presents the classical symmetrization and randomization tricks and decoupling inequality for degenerated weighted $U$-processes and the degenerated part $\tilde{U}_w(r)$. Then, some useful inequalities including the moment inequality are proved for weighted Rademacher chaos. We mainly use these inequalities to bound $U_w(r)$ and $\tilde{U}_w(r)$. Section D provides technical proofs omitted from the main track and the appendix. For the sake of completeness, we present the Khinchine inequality and the metric entropy inequality for Rademacher chaos in Section E.

A Janson’s Decomposition

There are literatures that use Janson’s decomposition derived from (Janson 2004) to study the problem learning from networked examples (Usunier, Amini, and Gallinari 2006, Biau and Bleakley 2006, Ralaivola, Szafranski, and Stempfel 2009, Ralaivola and Amini 2015). They usually model the networked data with the line graph of $G$.

Definition 6 (line graph). Let $G = (V, E)$ be a data graph we consider in the main track. We define the line graph of $G$ as a graph $D_G = (D_V, D_E)$, in which $D_V = E$ and $(i, j) \in D_E$ if and only if $i \cap e_j \neq \emptyset$ in $G$.

This framework differs from our setting and detains less information from the data graph, as argued in Section 2. One can analyze this framework by the fractional coloring and the Janson’s decomposition.

Definition 7 (fractional coloring). Let $D_G = (D_V, D_E)$ be a graph, $C = \{C_j, q_j\}_{j \in \{1, \ldots, r\}}$, for some positive integer $r$, with $C_j \subseteq D_V$ and $q_j \in [0, 1]$ is an fractional coloring of $D_G$ if

- $\forall j, C_j$ is an independent set, i.e., there is no connection between vertices in $C_j$.
- it is an exact cover of $G$: $\forall v \in D_V, \sum_{j: v \in C_j} q_j = 1$.

The weight $W(C)$ of $C$ is given by $W(C) = \sum_{j=1}^{r} q_j$ and the minimum weight $\chi^{*}(D_G) = \min_{C} W(C)$ over the set of all fractional colorings is the fractional chromatic number of $D_G$. By the Janson’s decomposition, one splits all examples into several independent sets according to a fractional coloring of $D_G$ and then analyze each set using the standard method for i.i.d. examples.

In particular, Theorem 3 (also see Usunier, Amini, and Gallinari 2006, Theorem 2) provides a variation of MCDiarmid’s theorem using the Janson’s decomposition. For simplicity, let $S = \{z_i\}_{i=1}^{m}$ be the training examples drawn from $Z^m$ and $S_C$ be the examples included in $C_j$. Also, let $k^j_i$ be the index of the $i$-th example in $C_j$ in the training set $S$.

Theorem 3. Using the notations defined above, let $C = \{C_j, q_j\}_{j=1}^{r}$ be a fractional coloring of $D_G$. Let $f: \mathbb{R}^m \rightarrow \mathbb{R}$ such that

- there exist $J$ functions $Z^{[C]}_i \rightarrow \mathbb{R}$ which satisfy $\forall S \in Z^m, f(S) = \sum_{j=1}^{J} q_j f_j(S_C)$.
- there exist $c_1, \ldots, c_m \in \mathbb{R}_+$ such that $\forall j \in C$, $S^k_C$ such that $S^k_C$ and $S^k_C$ differ only in the $k$-th example, $|f_j(S_C) - f_j(S^k_C)| \leq c_k j$.

Then, we have

$$P[f(S) - \mathbb{E}[f(S)] \geq \epsilon] \leq \exp \left(-\frac{2\epsilon^2}{\chi^*(D_G) \sum_{i=1}^{m} c_i^2} \right).$$

A.1 Weighted ERM

Using the above theorem, we show that the risk bounds for weighted ERM derived from the Janson’s decomposition cannot improve the results of (Usunier, Amini, and Gallinari 2006), as shown in the following theorem.

Theorem 4. Consider a minimizer $r_w$ of the weighted empirical risk $L_w$ over a class $R$. For all $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have

$$L(r) - L_w(r) \leq 9\mathbb{R}^{(w)}(R, S) + \sqrt{\chi^*(D_G) \log(1/\delta)} \frac{\|w\|_2}{\|w\|_1},$$

where

$$\mathbb{R}^{(w)}(R, S) = \frac{2}{N} \mathbb{E}_\sigma \left[ \sum_{j=1}^{J} \left( \sup_{r \in R} \sum_{c_j \in C_j} w_k_j r(z_k_j) \right) \right].$$

is the weighted empirical fractional Rademacher complexity of $R$ with respect to $D_G$.

In this setting, although the weights $w$ are no limit to the fractional matching, the risk bounds cannot improve the results of (Usunier, Amini, and Gallinari 2006), which is of the order $\sqrt{\chi^*(G)/m}$, as $\|w\|_2/\|w\|_1 \leq 1/\sqrt{m}$.

A.2 “Low-noise” Condition

If we considering the complete graph with equal weighting scheme, by the Janson’s decomposition, the empirical risk $L_m(L)$ can be represented as an average of sums of i.i.d. r.v.'s

$$\frac{1}{m!} \sum_{\pi \in \mathbb{P}_m} \frac{1}{m/2} \sum_{i=1}^{[m/2]} \ell(r, X_{\pi(i)}, X_{\pi(i)+[m/2]}, Y_{\pi(i), \pi(i)+[m/2]})$$

where the sum is taken over all permutations of $G_m$, the symmetric group of order $m$, and $[u]$ denotes the integer part of any $u \in \mathbb{R}$. From (Biau and Bleakley 2006), the bounds for excess risk of (15) are of the order $O(1/\sqrt{m})$. Moreover, by the result in (Ralaivola and Amini 2015), tighter risk bounds may be obtained under the following assumption which can lead to “low-noise” condition (Tsybakov 2004, Massart and Nédélec 2006).

Assumption 2. There exists $C > 0$ and $\theta \in [0, 1]$ such that for all $\epsilon > 0$,

$$P \left[ \left| \eta(X_1, X_2) - \frac{1}{2} \right| \leq \epsilon \right] \leq Ce^{\theta/(1-\theta)}.$$
Theorem 5. Suppose that one observes independent variable $X_i$, taking their values in some measurable space $X$. Think of $X$ as being the set of all measurable functions from $X$ to $\{0, 1\}$. Then we consider some loss function

$$\gamma : R \times \mathcal{Z} \to [0, 1].$$

Basically one can consider some set $\mathcal{R}$, which is known to contain the Bayes classifier $r^*$ that achieves the best (smallest) expected loss $P[\gamma(r, \cdot)]$ when $r$ varies in $\mathcal{R}$. The relative expected loss $\bar{\ell}$ is defined by

$$\bar{\ell}(r^*, r) = P[\gamma(r, \cdot) - \gamma(r^*, \cdot)], \forall r \in \mathcal{R} \quad (18)$$

Since the empirical process on i.i.d. examples split from the excess risk is with non-negative weights on all examples, we define the weighted loss as

$$\gamma_w(r) = \frac{1}{\|w\|_1} \sum_{i=1}^n w_i \gamma(r, \xi_i) \quad (19)$$

where $0 \leq w_i \leq 1$ for all $i = 1, \ldots, n$ and $\|w\|_1 = \sum_{i=1}^n w_i > 0$. Weighted ERM approach aims to find a minimizer of the weighted empirical loss $\bar{\ell}$ in the hypothesis set $R \subset \mathcal{R}$ to approximate $r^*$.

We introduce the weighted centered empirical process $\overline{\gamma}_w$ defined by

$$\overline{\gamma}_w(r) = \gamma_w(r) - P[\gamma(r, \cdot)] \quad (20)$$

In addition to the relative expected loss function $\bar{\ell}$, we shall need another way to measure the closeness between the elements of $R$. Let $d$ be some pseudo-distance on $\mathcal{R} \times \mathcal{R}$ such that

$$\text{Var}[\gamma(r, \cdot) - \gamma(r^*, \cdot)] \leq d^2(r, r^*), \forall r \in \mathcal{R}. \quad (21)$$

A tighter risk bound for weighted ERM is derived from Theorem 5 which combines two different moduli of uniform continuity: the stochastic modulus of uniform continuity of $\overline{\gamma}_w$ over $R$ with respect to $d$ and the modulus of uniform continuity of $d$ with respect to $\bar{\ell}$.

Next, we need to specify some mild regularity conditions functions that we shall assume to be verified by the moduli of continuity involved in the following result.

Definition 8. We denote by $D$ the class of nondecreasing and continuous functions $\psi$ from $\mathbb{R}_+$ to $\mathbb{R}_+$ such that $x \to \psi(x)/x$ is nonincreasing on $(0, +\infty)$ and $\psi(1) \geq 1$.

In order to avoid measurability problems, we need to consider some separability condition on $R$. The following one will be convenient.

Assumption 3. There exists some countable subset $R'$ of $R$ such that, for every $r \in R$, there exists some sequence $\{r_k\}$ of elements of $R'$ such that, for every $\xi \in \mathcal{Z}$, $\gamma(r_k, \xi, r)$ tends to $\gamma(r, \xi)$ as $k$ tends to infinity.

Assumption 4. There exists some countable subset $R''$ of $R$ such that, for every $r \in R$, there exists some sequence $\{r_k\}$ of elements of $R''$ such that, for every $\xi \in \mathcal{Z}$, $\gamma(r_k, \xi, \cdot)$ tends to $\gamma(r, \xi)$ as $k$ tends to infinity.

A special upper bound for the relative expected loss of any empirical risk minimizer on some given model $R$ will depend on the bias term $\ell(r^*, R) = \inf_{r \in R} \bar{\ell}(r^*, r)$ and the fluctuations of the empirical process $\overline{\gamma}_w$ on $R$. As a matter of fact, we shall consider some slightly more general estimators. Namely, given some nonnegative number $\rho$, we consider some $\rho$-empirical risk minimizer, that is, any estimator $r$ taking its values in $R$ such that $\gamma_w(r) \leq \rho + \inf_{r \in R} \gamma_w(r).$
Then there exists an absolute constant $K$.

**Theorem 6** (risk bound for weighted ERM). Let $\gamma$ be a loss function such that $r^* \gamma$ minimizes $P[\gamma(\cdot, \cdot)]$ when $r$ varies in $\mathcal{R}$. Let $\phi$ and $\psi$ belong to the class of functions $D$ defined above and let $R$ be a subset of $\mathcal{R}$ satisfying the separability Assumption 3. Assume that, on the one hand,

$$d(r^*, r) \leq \sqrt{\frac{1}{\|w\|_2^2} \phi(\sqrt{\ell(r^*, r)}}, \forall r \in \mathcal{R},$$

and that, on the other hand, one has, for every $r \in R'$,

$$\sqrt{\frac{1}{\|w\|_1} E \left[ \sup_{r' \in R', \frac{\|w\|_1}{\|w\|_2} d(r, r') \leq \sigma} \left( \frac{1}{\|w\|_1} \sup_{r' \in R'} \left( \sqrt{\ell(r', r)} \right) \right) \right] \leq \phi(\sigma)$$

for every positive $\sigma$ such that $\phi(\sigma) \leq \sqrt{\|w\|_1} \sigma^2$, where $R'$ is given by Assumption 3. Let $\epsilon_*$ be the unique positive solution of the equation

$$\sqrt{\frac{1}{\|w\|_1} \epsilon_*^2 = \phi(\epsilon_*)}.$$

Then there exists an absolute constant $K$ such that, for every $y \geq 1$, the following inequality holds:

$$P \left[ \ell(r^*, r) > 2\rho + 2\ell(r^*, R) + Ky\epsilon_*^2 \right] \leq e^{-y}.$$

**B.3 Maximal Inequality for Weighted Empirical Processes**

Next, we present the maximal inequality involved in covering number for weighted empirical processes. Let us fix some notation. We consider i.i.d. random variables $\xi_1, \ldots, \xi_n$ with values in some measurable space $\mathcal{Z}$ and common distribution $P$. For any $P$-integrable function $f$ on $\mathcal{Z}$, we define $P_w(f) = \frac{1}{\|w\|_1} \sum_{i=1}^n f(\xi_i)$ and $v_w(f) = P_w(f) - P(f)$ where $0 \leq w_i \leq 1$ for all $i = 1, \ldots, n$ and $\|w\|_1 = \sum_{i=1}^n w_i > 0$. Given a collection $\mathcal{F}$ of $P$-integrable functions $f$, our purpose is to control the expectation of $\sup_{f \in \mathcal{F}} v_w(f)$ or $\sup_{f \in \mathcal{F}} -v_w(f)$.

**Lemma 4.** Let $\mathcal{F}$ be a countable collection of measurable functions such that $f \in [0, 1]$ for every $f \in \mathcal{F}$, and let $f_0$ be a measurable function such that $f_0 \in [0, 1]$. Let $\sigma$ be a positive number such that $P[\|f - f_0]\| \leq \sigma^2$ for every $f \in \mathcal{F}$. Then, setting

$$\varphi(\sigma) = \int_0^\sigma (\log N_\infty(\mathcal{F}, \epsilon^2))^{1/2} d\epsilon,$$

the following inequality is available:

$$\sqrt{\|w\|_1} \max_{f \in \mathcal{F}} (E[\sup_{f \in \mathcal{F}} v_w(f_0)], E[\sup_{f \in \mathcal{F}} v_w(f - f_0)]) \leq 12\varphi(\sigma),$$

provided that $4\varphi(\sigma) \leq \sigma^2 \sqrt{\|w\|_1}$.

**C Inequalities for $U_w(r)$ and $\tilde{U}_w(r)$**

In this section, we first show the classical symmetrization and randomization tricks for the degenerated weighted $U$-statistics $U_w(r)$ and the degenerated part $\tilde{U}_w(r)$. Then we establish general exponential inequalities for weighted Rademacher chaos. This result is generalized from (Clémencçon, Lugosi, and Vayatis 2008) based on moment inequalities obtained for empirical processes and Rademacher chaos in (Boucheron et al. 2005). With this moment inequality, we prove the inequality for weighted Rademacher chaos, which involves the $L_\infty$ covering number of the hypothesis set.

**Lemma 5** (decoupling and undecoupling). Let $(X'_i)_{i=1}^n$ be an independent copy of the sequence $(X_i)_{i=1}^n$. Then, for all $q \geq 1$, we have:

$$E \left[ \sup_{f_{i,j} \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f_{i,j}(X_i, X_j) \right|^q \right] \leq 4^q E \left[ \sup_{f_{i,j} \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f_{i,j}(X'_i, X'_j) \right|^q \right] \leq 4^q E \left[ \sup_{f_{i,j} \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f_{i,j}(X_i, X_j) \right|^q \right].$$

If the functions $f_{i,j}$ are symmetric in the sense that for all $X_i, X_j$, $f_{i,j}(X_i, X_j) = f_{j,i}(X_j, X_i)$ and $(w_{i,j})_{i,j} \in \mathcal{E}$ is symmetric, then the inequality can be reversed, that is,

$$E \left[ \sup_{f_{i,j} \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f_{i,j}(X_i, X'_j) \right|^q \right] \leq 4^q E \left[ \sup_{f_{i,j} \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f_{i,j}(X'_i, X_j) \right|^q \right].$$

**Lemma 6** (randomization). Let $(\sigma_i)_{i=1}^n$ and $(\sigma'_i)_{i=1}^n$ be two independent sequences of i.i.d. Rademacher variables, independent from the $(X_i, X'_i)$’s. If $f$ is degenerated, we have for all $q \geq 1$,

$$E \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i,j} w_{i,j} f(X_i, X'_j) \right|^q \right] \leq 4^q E \left[ \sup_{f \in \mathcal{F}} \left| \sum_{i,j} \sigma_i \sigma'_j w_{i,j} f(X_i, X'_j) \right|^q \right].$$

**Lemma 7**. Let $(X'_i)_{i=1}^n$ be an independent copy of the sequence $(X_i)_{i=1}^n$. Consider random variables valued in $\{0, 1\}$, $(\tilde{Y}_{i,j})_{i,j} \in \mathcal{J}$, conditionally independent given the $X_i$’s and the $X'_i$’s and such that $P[Y_{i,j} = 1 | X_i, X'_j] = \eta(X_i, X'_j)$. We have for all $q \geq 1$,

$$E \left[ \sup_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{J}} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j}) \right]^q \leq 4^q E \left[ \sup_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{J}} w_{i,j} \tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) \right]^q \leq 4^q E \left[ \sup_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{J}} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j}) \right]^q$$

and the inequality can be reversed,

$$E \left[ \sup_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{J}} w_{i,j} \tilde{h}_r(X_i, X_j, \tilde{Y}_{i,j}) \right]^q \leq 4^q E \left[ \sup_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{J}} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j}) \right]^q$$
Lemma 8. Let \((\sigma_i)_{i=1}^{n}\) and \((\sigma'_i)_{i=1}^{n}\) be two independent sequences of i.i.d. Rademacher variables, independent from the \((X_i, X'_i, Y_{i,j}, \tilde{Y}_{i,j})\)’s. Then, we have for all \(q \geq 1\),

\[
\mathbb{E}\left[\sup_{r \in R} \left| \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X'_i, \tilde{Y}_{i,j}) \right|^q \right] \\
\quad \leq 4^q \mathbb{E}\left[\sup_{r \in R} \sum_{(i,j) \in E} \sigma_i \sigma'_j w_{i,j} \hat{h}_r(X_i, X'_i, \tilde{Y}_{i,j}) \right]^q
\]

Lemma 5 and 6 are applications of Theorem 3.1.1 and Theorem 3.5.3 of (De la Penya and Giné 2012) respectively, providing decoupling and randomization inequalities for degenerated weighted \(U\)-statistics of order 2. Lemma 7 and 8 are the variants of Lemma 5 and 6 respectively, which is suitable for the degenerated part \(\tilde{U}_w(r)\).

Theorem 7 (moment inequality). Let \(X, X_1, \ldots, X_n\) be i.i.d. random variables and let \(F\) be a class of kernels. Consider a weighted Rademacher chaos \(Z_\sigma\) of order 2 on the graph \(G = (V, E)\) indexed by \(F\),

\[
Z = \sup_{f \in F} \left| \sum_{(i,j) \in E} w_{i,j} f(X_i, X_j) \right|
\]

where \(\mathbb{E}[f(X, x)] = 0\) for all \(x \in \mathcal{X}, f \in F\). Assume also for all \(x, x' \in \mathcal{X}\), \(f(x, x') = f(x', x)\) (symmetric) and \(\sup_{f \in F} \|f\|_\infty = F\). Let \((\sigma_i)_{i=1}^{n}\) be i.i.d. Rademacher random variables and introduce the random variables

\[
Z_\sigma = \sup_{f \in F} \left| \sum_{(i,j) \in E} w_{i,j} \sigma_i \sigma_j f(X_i, X_j) \right|
\]

\[
U_\sigma = \sup_{f \in F} \sup_{|\alpha|_2 \leq 1} \left| \sum_{(i,j) \in E} w_{i,j} \sigma_i \alpha_j f(X_i, X_j) \right|
\]

\[
M = \sup_{f \in F, k = 1, \ldots, n} \left| \sum_{i:(i,k) \in E} w_{i,k} \sigma_i f(X_i, X_k) \right|
\]

Then exists a universal constant \(C\) such that for all \(n\) and \(t > 0\),

\[
P[Z \geq CE[Z_\sigma] + t] \\
\quad \leq \exp \left( -\frac{1}{C} \min \left( \frac{t}{\mathbb{E}[U_\sigma]}, \frac{t}{\mathbb{E}[M] + \mathbb{E}[w_i]} \right) \right)
\]

where \(\|w_i\|_\infty = \max_i \sqrt{\sum_{j:(i,j) \in E} w_{i,j}^2}\).

If the hypothesis set \(F\) is a subset of \(\mathcal{L}_\infty(\mathcal{X}^2)\) (upper bounds on the uniform covering number with \(\mathcal{L}_\infty\) metric can be calculated (Cucker and Zhou 2007)), we show \(\mathbb{E}[Z_\sigma], \mathbb{E}[U_\sigma]\) and \(\mathbb{E}[M]\) can be bounded by \(N_\infty(F, \epsilon)\) since all these Rademacher random variables satisfy the Khinchine inequality (see Section E). Following the metric entropy inequality for Khinchine-type processes (see Section E), it is easy to get the following Corollary.

Corollary 1. With the same setting of Theorem 7, if \(F \subset \mathcal{L}_\infty(\mathcal{X}^2)\), we have for any \(\delta < 1/e\),

\[
P[Z \leq \kappa] \geq 1 - \delta
\]

where

\[
\kappa \leq C \left( \|w\|_2 \int_0^{2^F} \log N_\infty(F, \epsilon) \, d\epsilon \right) + \max \left( \|w\|_2 \log(1/\delta) \int_0^{2^F} \sqrt{\log N_\infty(F, \epsilon)} \, d\epsilon, \right.
\]

\[
\left( \log(1/\delta) \right)^{3/2} \|w\|_{\max, \infty} \left( \log(1/\delta)^2 \|w\|_\infty \right)
\]

with a universal constant \(C\).

D. Technical Proofs

D.1 Proofs Omitted in Section 4

Proof of Lemma 1. Since \(R \subset \mathcal{L}_\infty(\mathcal{X}^2)\) (Assumption 1), by Corollary 1, the weighted degenerated \(U\)-process \(\sup_{r \in R} |U_w(r)|\) can be bounded by the \(L_\infty\) covering number of \(R\), that is, for any \(\delta \in (0, 1/e)\), we have

\[
P[\sup_{r \in R} |U_w(r)| \leq \kappa] \geq 1 - \delta
\]

where

\[
\kappa \leq C_1 \left( \|w\|_1 \int_0^{1/e} \log N_\infty(R, \epsilon) \, d\epsilon \right) + \max \left( \|w\|_2 \log(1/\delta) \int_0^{1/e} \sqrt{\log N_\infty(R, \epsilon)} \, d\epsilon, \right.
\]

\[
\left( \log(1/\delta) \right)^{3/2} \|w\|_{\max, \infty} \left( \log(1/\delta)^2 \|w\|_\infty \right)
\]

with a universal constant \(C_1 < \infty\). Then the first inequality for \(\sup_{r \in R} |U_w(r)|\) follows the fact that \(R\) satisfies Assumption 1. Similarly, by Lemma 7 and Lemma 8, we can convert the moment of \(\sup_{r \in R} |\tilde{U}_w(r)|\) to the moment of Rademacher chaos

\[
4^q \mathbb{E}\left[\sup_{r \in R} \left| \sum_{(i,j) \in E} \sigma_i \sigma'_j w_{i,j} \hat{h}_r(X_i, X'_i, \tilde{Y}_{i,j}) \right|^q \right]
\]

which can be handled by the by-products of Theorem 7. More specifically, using (46) and (47) combined with the arguments in Corollary 1 and Assumption 1 will gives the second inequality for \(\sup_{r \in R} |\tilde{U}_w(r)|\).

Proof of Lemma 2. For any function \(r \in \mathcal{R}\), observe first that

\[
\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1] \\
= \mathbb{E}[\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1, X_2] \mid X_1] \\
= \mathbb{E}[1 - 2\eta(X_1, X_2) \mathbb{I}_{r(X_1, X_2) \neq r^*(X_1, X_2)} \mid X_1]
\]

Then observing that

\[
|1 - 2\eta(X_1, X_2)|^2 \leq |1 - 2\eta(X_1, X_2)|
\]

almost sure, and combining with Jensen inequality, we have

\[
\mathbb{V}[\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1]]
\]
\[ \leq \mathbb{E}[(\mathbb{E}[e_r(X_1, X_2, Y_1, Y_2) \mid X_1])^2] \leq \mathbb{E}[\|1 - 2\eta(X_1, X_2)\|_{\mathbb{I}_{r(X_1, X_2)\neq r^*(X_1, X_2)}}] = \Lambda(r). \]

**Proof of Lemma 3.** First, we introduce some notations of weighted ERM of the i.i.d. case. We denote \( \{\bar{w}_i = \sum_{j(i,j) \in J} w_{i,j} : i = 1, \ldots, n\} \) the weights on vertices and introduce the “loss function” \( \gamma(r, X) = 2\beta_{ij}(X) + \Lambda(r) \) and the weighted empirical loss of vertices

\[ \gamma_\bar{w}(r) = \frac{1}{\|\bar{w}\|_1} \sum_{i=1}^n \bar{w}_i \gamma(r, X_i) = T_\bar{w}(r). \]

Define centered empirical process

\[ \overline{\gamma_\bar{w}}(r) = \frac{1}{\|\bar{w}\|_1} \sum_{i=1}^n \bar{w}_i (\gamma(r, X_i) - \Lambda(r)) \]

and the pseudo-distance

\[ d(r, r') = \frac{\sqrt{\|\bar{w}\|_1}}{\sqrt{\|\bar{w}\|_2}} \sqrt{\mathbb{E}[(\gamma(r, X) - \gamma(r', X))^2]} \]

for every \( r, r' \in R \). Let \( \phi \) be

\[ \phi(\sigma) = 12 \int_0^\sigma (\log N_{\infty} (R, \epsilon^2))^{1/2} d\epsilon. \tag{30} \]

From the definition of “loss function” \( \gamma \), we have the excess risk of \( r \)

\[ \bar{\ell}(r, r) = \Lambda(r) - \Lambda(r^*) = \Lambda(r). \]

According to Lemma 1, as \( \Lambda^2(r) \leq \Lambda(r) \), we have for every \( r \in R \),

\[ d(r, r^*) \leq \sqrt{\|\bar{w}\|_1 \sqrt{5 \overline{\ell}(r, r^*)}} \]

which implies that the modulus of continuity \( \psi \) can be taken as

\[ \psi(\epsilon) = \sqrt{5}\epsilon. \tag{31} \]

Then from Lemma 4, we have

\[ \sqrt{\|\bar{w}\|_1 E} \left[ \sup_{r' \in R, \|\bar{w}\|_2 \leq \sigma} |\overline{\gamma_\bar{w}}(r) - \overline{\gamma_\bar{w}}(r')| \right] \leq \phi(\sigma). \]

provided that \( \phi(\sigma)/3 \leq \sqrt{\|\bar{w}\|_1} \sigma^2 \). It remains to bound the excess risk of \( r_\bar{w} \) by the tight bounds for weighted ERM on i.i.d. examples by Theorem 6. For any \( \delta \in (0, 1) \), we have with probability at least \( 1 - \delta \),

\[ L(r_\bar{w}) - L^* \leq 2(\inf_{r \in R} L(r) - L^*) + 2\rho + K(\log(1/\delta))\epsilon^2 \]

where \( C \) is a universal constant and \( \epsilon_\delta \) is the unique positive solution of the equation

\[ \sqrt{\|\bar{w}\|_1} \epsilon_\delta^2 = \phi(\psi(\epsilon_\delta)). \]

When \( R \) satisfies Assumption 1, there exists a universal constant \( C' \) such that

\[ \epsilon_\delta^2 \leq C' K^{1/(1+\beta)} \left( \frac{1}{(1 - \beta)^2\|\bar{w}\|_1} \right)^{1/(1+\beta)} \]

which completes the proof. \( \square \)

**D.2 Proof omitted in Section A**

**Proof of Theorem 4.** We write, for all \( r \in R \),

\[ L(r) - L_{\bar{w}}(r) \leq \sup_{r \in R} \left[ \mathbb{E}[\ell(r, z)] - \frac{1}{\|\bar{w}\|_1} \sum_{i=1}^m w_i \ell(r, z_i) \right] \]

\[ \leq \sum_{j=1}^J \left[ \sup_{r \in R} \sum_{i \in J_j} \frac{w_{k_j}}{\|\bar{w}\|_1} (\mathbb{E}[\ell(r, z)] - \ell(r, z_{k_j})) \right]. \tag{33} \]

Now, consider, for each \( j \),

\[ f_j(S_{C_j}) = \sup_{r \in R} \sum_{i \in J_j} \frac{w_{k_j}}{\|\bar{w}\|_1} (\mathbb{E}[\ell(r, z)] - \ell(r, z_{k_j})). \]

Let \( f \) be defined by, for all training set \( S \), \( f(S) = \sum_{j=1}^J p_j f_j(S_{C_j}) \), then \( f \) satisfies the conditions of Theorem 3 with, for any \( i \in \{1, \ldots, m\} \), \( \beta_i \leq w_i/\|\bar{w}\|_1 \). Therefore, we can claim that, with probability at least \( 1 - \delta \),

\[ L(r) - L_{\bar{w}}(r) \leq \sup_{r \in R} \left[ \mathbb{E}[\ell(r, z)] - \frac{1}{\|\bar{w}\|_1} \sum_{i=1}^m w_i \ell(r, z_i) \right] \]

\[ + \sqrt{\chi^2 (D_G) \log(1/\delta)} \frac{\|\bar{w}\|_2}{\|\bar{w}\|_1} \]

Then, using the standard symmetrization technique (see Usunier, Amini, and Gallinari 2006, Theorem 4), one can bound the first item in the right hand side by \( \mathfrak{R}_{\bar{w}}(R, S) \) which completes the proof. \( \square \)

**D.3 Proofs omitted in Section B**

**Proof of Theorem 5.** We use an auxiliary random variable

\[ \tilde{Z} = \frac{\|\bar{w}\|_1}{2b} Z = \sup_{f \in \mathcal{F}} \frac{1}{2b} \sum_{i=1}^n w_i (f(X_i) - \mathbb{E}[f(X_i)]). \]

We denote by \( f_k \) a function such that

\[ f_k = \frac{1}{2b} \sup_{f \in \mathcal{F}} \sum_{i \neq k} w_i (f(X_i) - \mathbb{E}[f(X_i)]). \]

We introduce following auxiliary random variables for \( k = 1, \ldots, n \),

\[ Z_k = \frac{1}{2b} \sup_{f \in \mathcal{F}} \sum_{i \neq k} w_i (f(X_i) - \mathbb{E}[f(X_i)]). \]

and

\[ Z'_k = \frac{1}{2b} w_k (f(X_k) - \mathbb{E}[f(X_k)]). \]

Denoting by \( f_0 \) the function achieving the maximum in \( Z \), we have

\[ \tilde{Z} - Z_k \leq \frac{1}{2b} w_k (f_0(X_k) - \mathbb{E}[f_0(X_k)]) \leq 1 \text{ a.s.,} \]

\[ \tilde{Z} - Z_k - Z'_k \geq 0 \]

and

\[ \mathbb{E}[Z'_k] = 0. \]
The first inequality is derived from \( w_i \leq 1 \) and 
\[ \sup_{f \in \mathcal{F}, X \in X} f(X) - \mathbb{E}[f(X)] \leq 2b. \] Also, we have 
\[ (n-1)\bar{Z} = \sum_{k=1}^{n} \frac{1}{2b} \sum_{i \neq k} w_i (f_0(X_i) - \mathbb{E}[f_0(X_i)]) \leq \sum_{k=1}^{n} Z_k, \]
and 
\[ \sum_{k=1}^{n} \mathbb{E}_n[Z_k^2] = \frac{1}{2b} \sum_{k=1}^{n} \mathbb{E}[w_i^2 (f_k(X_k) - \mathbb{E}[f_k(X_k)])^2] \leq \frac{1}{4b^2} \|w\|_2^2 \text{sup}_{f \in \mathcal{F}} \text{Var}[f(X)] \leq \frac{1}{4b^2} \|w\|_2^2 \sigma^2. \]
where \( \sigma^2 \geq \sup_{f \in \mathcal{F}} \text{Var}[f(X)] \). Notice that we use the fact the \( X_i \) have identical distribution. Applying Theorem 1 of (Bousquet 2002) with \( v = 2\sqrt{\bar{Z}} + \frac{\|w\|_2^2}{4b^2} \sigma^2 \) will give 
\[ P[\bar{Z} - \mathbb{E}[\bar{Z}] \geq \sqrt{2v} + \frac{x}{3}] \leq e^{-x}, \]
and then 
\[ P\left[ \frac{\|w\|_1}{2b} (Z - \mathbb{E}[Z]) \geq \sqrt{2x \left( \frac{\|w\|_1 \mathbb{E}[Z]}{b} + \frac{\|w\|_2^2}{4b^2} \sigma^2 \right)} + \frac{x}{3} \right] \leq e^{-x}, \]
which proves the inequality. \( \square \)

**Proof of Theorem 6.** Since \( R \) satisfies Condition 3, we notice that, by dominated convergence, for every \( r \in R \), considering the sequence \( \{r_k\} \) provided by Condition 3, one has \( P[\gamma(r, r_k)] \) that tends to \( P[\gamma(r, r)] \) as \( k \) tends to infinity. Denote the bias term of loss \( \bar{\ell}(r^*, R) = \inf_{r \in R} \ell(r^*, r) \). Hence, \( \ell(r^*, R) = \bar{\ell}(r^*, r, R) \), which implies that there exists some point \( \pi(r^*) \) (which may depend on \( \epsilon_* \)) such that \( \pi(r^*) \in R^c \) and 
\[ \bar{\ell}(r^*, \pi(r^*)) \leq \bar{\ell}(r^*, R) + \epsilon_*^2. \]

We start from the identity 
\[ \bar{\ell}(r^*, \hat{r}) = \bar{\ell}(r^*, \pi(r^*)) + \gamma_w(\hat{r}) - \gamma_w(\pi(r^*)) + \bar{\ell}(r^*, \hat{r}) \]
which, by definition of \( \hat{r} \), implies that 
\[ \tilde{\ell}(r^*, \hat{r}) \leq \rho + \bar{\ell}(r, \pi(r^*)) + \bar{\ell}(r^*, \hat{r}) \]
Let \( x = \sqrt{K}\gamma_{\epsilon_*} \), where \( K' \) is a constant to be chosen later such that \( K' \geq 1 \) and 
\[ V_x = \sup_{r \in R} \frac{\tau_w(\pi(r^*)) - \tau_w(r)}{\ell(r^*, r) + \epsilon_*^2 + x^2 x^2}. \]
Then, 
\[ \tilde{\ell}(r^*, \hat{r}) \leq \rho + \bar{\ell}(r^*, \pi(r^*)) + V_x(\tilde{\ell}(r^*, \hat{r}) + x^2 + \epsilon_*^2) \]
and therefore, on the event \( V_x < 1/2 \), one has 
\[ \ell(r^*, \hat{r}) \leq 2(\rho + \ell(r^*, \pi(r^*))) + \epsilon_*^2 + x^2, \]
yielding 
\[ P[\ell(r^*, \hat{r}) \leq 2(\rho + \ell(r^*, \pi(r^*))) + 3\epsilon_*^2 + x^2] \leq P[V_x \geq 1/2]. \] \( (35) \)

Since \( \ell \) is bounded by 1, we may always assume \( x \) (and thus \( \epsilon_* \)) to be not larger than 1. Assuming that \( x \leq 1 \), it remains to control the variable \( V_x \) via Theorem 5. In order to use Theorem 5, we first remark that, by Condition 3, 
\[ V_x = \sup_{r \in R} \frac{\tau_w(\pi(r^*)) - \tau_w(r)}{\ell(r^*, r) + \epsilon_*^2 + x^2 \epsilon x^2} \]
which means that we indeed have to deal with a countably indexed empirical process. Note that the triangle inequality implies via (21), (34) and (22) that 
\[ (\text{Var}[\gamma(r, \cdot) - \gamma(\pi(r^*), \cdot)])^{\frac{1}{2}} \leq d(r^*, r) + d(r^*, \pi(r^*)) \]
\[ \leq 2\sqrt{\mathbb{E}[V_x \sqrt{\ell(r^*, r) + \epsilon_*^2 + x^2 \epsilon x^2}]} \]
\[ (36) \]
Since \( \gamma \) takes its values in \([0, 1]\), introducing the functions \( \psi_1 = \min(1, 2\psi) \) and, we derive from (36) that 
\[ \sup_{r \in R} \text{Var} \left[ \frac{\tau_w(\pi(r^*)) - \tau_w(\psi)}{\ell(r^*, r) + \epsilon_*^2 + x^2 \epsilon x^2} \right] \leq \sup_{\epsilon \geq 0} \frac{\mathbb{E}[\psi]^2}{\epsilon_0^2} \leq \frac{\mathbb{E}[\psi]^2}{\epsilon_0^2} \leq \frac{\mathbb{E}[\psi]^2}{\epsilon_0^2} \]
Now the monotonicity assumptions on \( \psi \) imply that either \( \psi(\epsilon) \leq \psi(\epsilon) \) if \( \epsilon \geq \epsilon \) or \( \psi(\epsilon) \leq \psi(x) \) if \( \epsilon \leq \epsilon \). Hence, one has in any case \( \psi(\epsilon)(\max(\epsilon, x)) \leq \psi(x)/x \), which finally yields 
\[ \sup_{r \in R} \text{Var} \left[ \frac{\tau_w(\pi(r^*)) - \tau_w(\psi)}{\ell(r^*, r) + \epsilon_*^2 + x^2 \epsilon x^2} \right] \leq \frac{\mathbb{E}[\psi]^2}{\epsilon_0^2} \]
\[ \mathbb{E}[V_x + 2\sqrt{\mathbb{E}[\psi]^2 x}\mathbb{E}[V_x]} \leq \frac{2y}{3\|w\|_1 x^2 \epsilon x^2} \]
\[ (37) \]
Now since \( \epsilon_* \) is assumed to be not larger than 1, one has \( \psi(\epsilon_* \geq \epsilon_1 \) and therefore, for every \( \sigma \geq \psi(\epsilon_1) \), the following inequality derives from the definition of \( \epsilon_* \) by monotonicity: 
\[ \frac{\phi(\sigma)}{\sigma^2} \leq \frac{\phi(\psi(\epsilon_*))}{\psi(\epsilon_1) \epsilon_*^2} \leq \frac{\phi(\psi(\epsilon_*))}{\epsilon_*^2} = \sqrt{\|w\|_1}. \]
Thus, (23) holds for every $\sigma \geq \psi(\epsilon_s)$. In order to control $E[V_2]$, we intend to use Lemma A.5 of (Massart and Nédélec 2006). For every $r \in R'$, we introduce $a^2(r) = \max(\ell(r^*, \pi(r^*)), \ell(r^*, r))$. Then by (34), $\ell(r^*, r) \leq a^2(r) \leq \ell(r^*, r) + \epsilon_s^2$. Hence, we have, on the one hand, that

$$E[V_2] \leq E[\sup_{r \in R'} \sqrt{w}(\pi(r^*)) - \sqrt{w}(r)].$$

and, on the other hand, that, for every $\epsilon \geq \epsilon_s$,

$$E[\sup_{r \in R', \ell(r^*, r) \leq \epsilon} \sqrt{w}(\pi(r^*)) - \sqrt{w}(r)] \leq E[\sup_{r \in R', \ell(r^*, r) \leq \epsilon} \sqrt{w}(\pi(r^*)) - \sqrt{w}(r)].$$

Now by (34) if there exists some $r \in R'$ such that $\ell(r^*, r) \leq \epsilon^2$, then $\ell(r^*, \pi(r^*)) \leq \epsilon^2 + \epsilon_s^2 \leq 2\epsilon^2$ and therefore, by assumption (22) and monotonicity of $\theta \rightarrow \psi(\theta)/\theta$, $d(\pi(r^*), r) \leq 2\sqrt{2} \|w\|_1 \psi(\sqrt{2} \epsilon) \leq 2\sqrt{2} \psi(\|w\|_1 \psi(\epsilon))$, then $\frac{\|w\|_1}{\sqrt{2}} d(\pi(r^*), r) \leq 2\sqrt{2} \psi(\epsilon)$. Thus, we derive from (23) that, for every $\epsilon \geq \epsilon_s$,

$$E[\sup_{r \in R', \ell(r^*, r) \leq \epsilon^2} \sqrt{w}(\pi(r^*)) - \sqrt{w}(r)] \leq \phi(2\sqrt{2} \psi(\epsilon))$$

and since $\theta \rightarrow \phi(2\sqrt{2} \psi(\epsilon))/\theta$ is nonincreasing, we can use Lemma A.5 of (Massart and Nédélec 2006) to get

$$E[V_2] \leq 4\phi(2\sqrt{2} \psi(x))/\sqrt{\|w\|_1 x^2},$$

and by monotonicity of $\theta \rightarrow \phi(\psi(\theta))/\theta$,

$$E[V_2] \leq 8\sqrt{2} \psi(\|x\|_1)/\sqrt{\|w\|_1 x^2}.$$ 

Thus, using the monotonicity of $\theta \rightarrow \phi(\psi(\theta))/\theta$, and the definition of $\epsilon_s$, we derive that

$$E[V_2] \leq \frac{8\sqrt{2} \psi(\psi(\epsilon_s))}{\sqrt{\|w\|_1 x^2}} \leq \frac{8\sqrt{2} \psi(\epsilon_s)}{\sqrt{K' y}} \leq \frac{8\sqrt{2} \psi(\epsilon_s)}{\sqrt{K'}} \tag{38}$$

provided that $x \geq \epsilon_s$, which holds since $K' \geq 1$. Now, the monotonicity of $\theta \rightarrow \psi(\theta)/\theta$ implies that $\epsilon_s^2 \psi(\epsilon_s) \leq \epsilon_s^2 \psi(\epsilon_s)$, but since $\phi(\psi(\theta)/\theta) \geq \phi(1)$ for every $\theta$ in $[0, 1]$, we derive from (24) and the monotonicity of $\psi$ and $\theta \rightarrow \phi(\psi(\theta))/\theta$ that

$$\frac{\psi(\epsilon_s)}{\epsilon_s^2} \leq \frac{\phi(\psi(\epsilon_s))}{\epsilon_s^2} \leq \frac{4\phi(\psi(\epsilon_s))}{\epsilon_s^2}$$

and, therefore, $x^2 \psi(x) \leq 4\|w\|_1 \epsilon_s^2$. Plugging this inequality together with (38) into (37) implies that, on the set $\Omega_y$, $V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \frac{2y(4\|w\|_1 \epsilon_s^2 + 2\sqrt{2} K')}{\|w\|_1 x^2} + \frac{2y}{3\|w\|_1 x^2}$.

It remains to replace $x^2$ by its value $K'y^2 \epsilon_s^2$ to derive that, on the set $\Omega_y$, the following inequality holds:

$$V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \frac{8(1 + 4\|w\|_1 \epsilon_s^2 \sqrt{K'})}{K'} + \frac{2}{3\|w\|_1 K' y^2}.$$

Taking into account that $\phi(\psi(\theta))/\theta \geq \phi(\min(1, \psi(\theta)))$ for every $\theta \in [0, 1]$, we deduce from the definition of $\epsilon_s$ that $\|w\|_1 \epsilon_s^2 \geq 1$ and, therefore, the preceding inequality becomes, on $\Omega_y$, $V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \frac{8(1 + 4/\sqrt{K'})}{K'} + \frac{2}{3K'}$.

Hence, choosing $K'$ as a large enough numerical constant guarantee that $V_x < 1/2$ on $\Omega_y$ and, therefore, (35) yields $P[\ell(r^*, r) \leq 2(\rho + \ell(r^*, \pi(r^*))) + 3\epsilon^2 + x^2] \leq P[\Omega_y^c] \leq e^{-y}$.

We get the required probability bound (6) by setting $K = K' + 3$. 

**Proof of Lemma 4.** We first perform the control of $E[\sup_{f \in F} v_w(f - f_0)]$. For simplicity, we denote $H_{\infty}(F, \epsilon) = \log N_{\infty}(F, \epsilon)$. For any integer $j$, we set $\sigma_j = \sigma_j^2$ and $H_j = H_{\infty}(F, \sigma_j^2)$. By definition of $H_j = H_{\infty}(F, \sigma_j^2)$, for any integer $j \geq 1$, we can define a mapping $\Pi_j$ from $F$ to some finite collection of functions such that

$$\log \# \{|\Pi_j F| \leq H_j \} \geq 0 \leq \Pi_j f \leq 1.$$ 

Noticing that since $u \rightarrow H_{\infty}(F, u^2)$ is nonincreasing, $H_1 \leq \sigma_1^{-2} \psi^2(\sigma)$, and under the condition $4\sigma r(x) \leq \sigma \psi(\|w\|_1)$, one has $H_1 \leq \sigma_1^2 \sigma \|w\|_1$. Thus, since $j \rightarrow H_j \sigma_j^{-2}$ increases to infinity, the set $\{j \geq 0 : H_j \leq \sigma_j^2 \|w\|_1 \}$ is nonvoid interval of the form $\{j \geq 0 : H_j \leq \sigma_j^2 \|w\|_1 \} = [0, J]$, with $J \geq 1$. For every $f \in F$, starting from the decomposition

$$-v_w(f) = \sum_{j=0}^{J-1} v_w(\Pi_j f) - v_w(\Pi_{j+1} f) + v_w(\Pi_j f) - v_w(f),$$

we derive, since $\Pi_j(f) \leq f$ and $P(f - \Pi_j(f)) \leq \sigma_j^2$, that

$$-v_w(f) = \sum_{j=0}^{J-1} v_w(\Pi_j f) - v_w(\Pi_{j+1} f) + \sigma_j^2.$$

Therefore, we can use the previous results to control $E[\sup_{f \in F} v_w(f - f_0)]$. For simplicity, we denote $H_{\infty}(F, \epsilon) = \log N_{\infty}(F, \epsilon)$. For any integer $j$, we set $\sigma_j = \sigma_j^2$ and $H_j = H_{\infty}(F, \sigma_j^2)$. By definition of $H_j = H_{\infty}(F, \sigma_j^2)$, for any integer $j \geq 1$, we can define a mapping $\Pi_j$ from $F$ to some finite collection of functions such that

$$\log \# \{|\Pi_j F| \leq H_j \} \geq 0 \leq \Pi_j f \leq 1.$$ 

Noticing that since $u \rightarrow H_{\infty}(F, u^2)$ is nonincreasing, $H_1 \leq \sigma_1^{-2} \psi^2(\sigma)$, and under the condition $4\sigma r(x) \leq \sigma \psi(\|w\|_1)$, one has $H_1 \leq \sigma_1^2 \sigma \|w\|_1$. Thus, since $j \rightarrow H_j \sigma_j^{-2}$ increases to infinity, the set $\{j \geq 0 : H_j \leq \sigma_j^2 \|w\|_1 \}$ is nonvoid interval of the form $\{j \geq 0 : H_j \leq \sigma_j^2 \|w\|_1 \} = [0, J]$, with $J \geq 1$. For every $f \in F$, starting from the decomposition

$$-v_w(f) = \sum_{j=0}^{J-1} v_w(\Pi_j f) - v_w(\Pi_{j+1} f) + v_w(\Pi_j f) - v_w(f),$$

we derive, since $\Pi_j(f) \leq f$ and $P(f - \Pi_j(f)) \leq \sigma_j^2$, that

$$-v_w(f) = \sum_{j=0}^{J-1} v_w(\Pi_j f) - v_w(\Pi_{j+1} f) + \sigma_j^2.$$
and, therefore,
\[ \mathbb{E}[\sup_{f \in \mathcal{F}} -v_\mathbf{w}(f)] \]
\[ \leq \sum_{j=0}^{J-1} \mathbb{E}[\sup_{f \in \mathcal{F}} v_\mathbf{w}(\Pi_j f) - v_\mathbf{w}(\Pi_j+1 f)] + \sigma_j^2. \]  
(42)

Now, it follows from (40) and (41) that, for every integer \( j \) and every \( f \in \mathcal{F} \), one has
\[ P[|\Pi_j f - \Pi_j+1 f|] \leq \sigma_j^2 + \sigma_{j+1}^2 = 5\sigma_{j+1}^2 \]
and, therefore, since \(|\Pi_j f - \Pi_j+1 f| \leq 1\),
\[ P[|\Pi_j f - \Pi_j+1 f|^2] \leq 5\sigma_{j+1}^2. \]
Moreover, (39) ensures that the number of functions of the form \( \Pi_j f - \Pi_j+1 f \) when varies in \( \mathcal{F} \) is not larger than \( \exp(H_j + H_{j+1}) \leq \exp(2H_{j+1}) \). Hence, we derive from the maximal inequality for random vectors (see Massart and Nédélec 2006, Lemma A.1) and the by-product of the proof of Bernstein’s inequality for the weighted sum of networked random variables (see Wang, Guo, and Ramon 2017, Lemma 16) that
\[ \sqrt{|\mathbf{w}|_1^2 \mathbb{E}[\sup_{f \in \mathcal{F}} |v_\mathbf{w}(\Pi_j f) - v_\mathbf{w}(\Pi_j+1 f)|]} \]
\[ \leq 2[\sigma_{j+1}^2 + 5H_{j+1} + \frac{1}{3\sqrt{|\mathbf{w}|_1}} H_{j+1}] \]
because \( w_i \leq 1, \forall i \in 1, \ldots, n \), and (42) becomes
\[ \sqrt{|\mathbf{w}|_1^2 \mathbb{E}[\sup_{f \in \mathcal{F}} -v_\mathbf{w}(f)]} \]
\[ \leq 2 \sum_{j=1}^{J} |\sigma_j| \sqrt{5H_j + \frac{1}{3\sqrt{|\mathbf{w}|_1}} H_j} + 4\sqrt{|\mathbf{w}|_1} \sigma_{J+1}. \]  
(43)

It follows from the definition of \( J \) that, on the one hand, for every \( j \leq J \),
\[ \frac{1}{3\sqrt{|\mathbf{w}|_1}} H_j \leq \frac{1}{3\sqrt{H_J}} \]
and, on the other hand,
\[ 4\sqrt{|\mathbf{w}|_1} \sigma_{J+1} \leq 4\sigma_{J+1} \sqrt{H_{J+1}}. \]
Hence, plugging these inequalities in (43) yields
\[ \sqrt{|\mathbf{w}|_1^2 \mathbb{E}[\sup_{f \in \mathcal{F}} -v_\mathbf{w}(f)]} \leq 6 \sum_{j=1}^{J+1} |\sigma_j| \sqrt{H_j}, \]
and the result follows. The control of \( \mathbb{E}[\sup_{f \in \mathcal{F}} v_\mathbf{w}(f - f_0)] \) can be performed analogously. \( \square \)

D.4 Proofs Omitted in Section C

**Proof of Lemma 7.** This Lemma is derived from Lemma 5, thus we can follow the similar arguments that can be found in (De la Peña and Giné 2012).

For any random variable \( X \), we denote by \( \mathcal{L}(X) \) its distribution. We denote by \( \Sigma \) (respectively \( \Sigma' \)) the sigma-filed generated by \{\( X_1, \ldots, X_n \)\} (respectively \{\( X'_1, \ldots, X'_n \)\}). Let \( \{Y_{i,j}^r\}_{(i,j) \in E} \) be Bernoulli random variables such that \( P[Y_{i,j}^r = 1 | \Sigma, \Sigma'] = \eta(X_i, X'_j) \). Let \( \{\sigma_i\}_{i=1}^n \) be independent Rademacher variables and define:
\[ Z_i = X_i \text{ if } \sigma_i = 1, \text{ and } X'_i \text{ otherwise}, \]
\[ Z'_i = X'_i \text{ if } \sigma_i = 1, \text{ and } X_i \text{ otherwise}. \]
Conditionally upon the \( X_i \) and \( X'_i \), the random vector \( (Z_i, Z'_i) \) takes the values \( (X_i, X'_i) \) or \( (X'_i, X_i) \), each with probability 1/2. In particular, we have:
\[ \mathcal{L}(X_1, \ldots, X_n, X'_1, \ldots, X'_n) = \mathcal{L}(Z_1, \ldots, Z_n, Z'_1, \ldots, Z'_n) \]  
(44)
and
\[ \mathcal{L}(X_1, \ldots, X_n) = \mathcal{L}(Z_1, \ldots, Z_n). \]  
(45)

Let \( \{\tilde{Y}_{i,j}\}_{(i,j) \in E} \) be Bernoulli random variables such that \( P[\tilde{Y}_{i,j} = 1 | \Sigma, \Sigma'] = \eta(X_i, X'_j) \) and define for \( (i,j) \in E \),
\[ \tilde{Y}_{i,j} \begin{cases} Y_{i,j} & \text{if } \sigma_i = 1 \text{ and } \sigma_j = -1 \\ Y'_{i,j} & \text{if } \sigma_i = -1 \text{ and } \sigma_j = 1 \\ Z_i & \text{if } \sigma_i = 1 \text{ and } \sigma_j = 1 \\ Z'_i & \text{if } \sigma_i = -1 \text{ and } \sigma_j = 1 \end{cases} \]
Notice that for all \( f \),
\[ \mathbb{E}_\sigma[\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j})] \]
\[ = \frac{1}{4} \left( \tilde{h}_r(X_i, X_j, Y_{i,j}) + \tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) + \tilde{h}_r(X_i, X'_j, Y'_{i,j}) + \tilde{h}_r(X'_i, X'_j, Y'_{i,j}) \right) \]
where \( \mathbb{E}_\sigma \) denotes the expectation taken with respect to \( \{\sigma_i\}_{i=1}^n \). Moreover, using
\[ \mathbb{E}[\tilde{h}_r(X'_i, X'_j, Y'_{i,j}) | \Sigma] = 0 \]
and (degenerated)
\[ \mathbb{E}[\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) | \Sigma] = 0 \]
\[ \mathbb{E}[\tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) | \Sigma] = 0 \]
we easily get
\[ \tilde{h}_r(X_i, X_j, Y_{i,j}) = 4\mathbb{E}[\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j}) | \Sigma] \]
For all \( q \geq 1 \), we therefore have
\[ \mathbb{E}[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j})^q] \]
\[ = \mathbb{E}[\sup_{r \in R} \sum_{(i,j) \in E} 4w_{i,j} \mathbb{E}[\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j}) | \Sigma]^q] \]
\[ \leq 4^q \mathbb{E}[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j})]^{q}] \]
derived from the facts that the supreme and \( |x|^p(p \geq 1) \) are convex functions and the Jansen inequality. According to (44) and the fact that the distribution of \( \tilde{Y}_{i,j} \) only depends on
which concludes the proof of (28).

By the symmetry of $\hat{h}_r$ in the sense that $\hat{h}_r(X_1, X_2, Y_{i,j}) = \hat{h}_r(X_2, X_1, Y_{i,j})$, we have

$$ E[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, Y_{i,j})] $$

which (29) follows.

**Proof of Lemma 8.** Re-using the notations used in the proof of Lemma 5, we further introduce $(X''_i)^n_{i=1}$, a copy of $(X'_i)^n_{i=1}$, independent from $\Sigma, \Sigma'$, and denote by $\Sigma''$ its sigma-field. Let $(\hat{Y}_{i,j}''')_{(i,j) \in E}$ Bernoulli random variables such that $P[\hat{Y}_{i,j}''' = 1 \mid \Sigma, \Sigma', \Sigma''] = \eta(X_i, X''_j)$. We now use classic randomization techniques and introduce our "ghost" sample:

$$ E[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, Y_{i,j}''')] $$

$$(\hat{h}_r \text{ is degenerated})$$

$$ = E[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, \hat{Y}_{i,j}''')] $$

(Jansen’s Inequality)

$$ \leq E[\sup_{r \in R} \sum_{j=1}^n \sum_{i: (i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, \hat{Y}_{i,j}''')] $$

Let $\sigma_{i,j}^{n_{i,j}}$ be independent Rademacher variables, independent of $\Sigma, \Sigma'$ and $\Sigma''$, then we have:

$$ E[\sup_{r \in R} \sum_{i,j} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, \hat{Y}_{i,j})] $$

$$ + E[\sup_{r \in R} \sum_{i,j} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, Y_{i,j}''')] $$

(Jensen’s Inequality and the convexity of $\sup_{r \in R} \cdot$)

$$ \leq E[\sup_{r \in R} \sum_{i,j} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, \hat{Y}_{i,j})] $$

$$ + 2E[\sup_{r \in R} \sum_{i,j} \sum_{(i,j) \in E} w_{i,j} \hat{h}_r(X_i, X_j, Y_{i,j}''')] $$

(According to (45))
and get
\[ \mathbb{E}[\sup_{r \in R} \sum_{(i,j) \in E} w_{i,j} \tilde{h}_r(X_i, X_j, \tilde{Y}_{i,j})] \]
\[ \leq 2^q \mathbb{E}[\sup_{r \in R} \sum_{j=1}^n \sigma_j \sum_{i \in (i,k) \in E} w_{i,j} \tilde{h}_r(X_i, X_j, \tilde{Y}_{i,j})] \]

Then repeating the same argument but for the \((X_i)_{i=1}^n\) will give the similar inequality. The desired inequality will follows putting these two inequalities together.

**Proof of Theorem 7.** By the decoupling, undecoupling and randomization techniques (see Lemma 5, Lemma 6), the symmetry and the degeneration of \( f \) and the symmetry of \((w_{i,j})_{(i,j) \in E} \), we have
\[ \mathbb{E}[\sup_f \sum_{(i,j) \in E} w_{i,j} f(X_i, X_j)] \]
\[ \leq 16^q \mathbb{E}[\sup_f \sum_{(i,j) \in E} w_{i,j} \sigma_i \sigma_j f(X_i, X_j)] \]
\[ \leq 64^q \mathbb{E}[\sup_{(i,j) \in E} \sum_{i,j} \sigma_i \sigma_j f(X_i, X_j)] \]

It means we can convert the moment of the original \( U \)-process to the moment of Rademacher chaos which can be handled by moment inequalities of (Boucheron et al. 2005).

In particular, for any \( q \geq 2 \),
\[ (\mathbb{E}_\sigma[Z_r^q])^{1/q} \leq \mathbb{E}_\sigma[Z_r] + (\mathbb{E}_\sigma[(Z_r - \mathbb{E}_\sigma[Z_r])^q])^{1/q} \]
\[ \leq \mathbb{E}_\sigma[Z_r] + 3 \sqrt{q} \mathbb{E}_\sigma U_r + 4q B \]

where \( B \) is defined below
\[ B = \sup_{\alpha, \alpha'} \sum_{i,j} |w_{i,j} \alpha_i \alpha'_j f(X_i, X_j)|. \]

The second inequality above follows by Theorem 14 (Boucheron et al. 2005).

Using the inequality \( (a + b + c)^q \leq 3^{q-1} (a^q + b^q + c^q) \) valid for \( q \geq 2 \), \( a, b, c > 0 \), we have
\[ \mathbb{E}_\sigma[Z_r^q] \leq 3^{q-1} (\mathbb{E}_\sigma[Z_r]^q + 3q \mathbb{E}_\sigma[U_r]^q + 4q^2 B^q). \]

It remains to derive suitable upper bounds for the expectation of the three terms on the right hand side.

**First term:** \( \mathbb{E}[\mathbb{E}_\sigma[Z_r]^q] \). Using the symmetrization trick, we have
\[ \mathbb{E}[\mathbb{E}_\sigma[Z_r]^q] \leq 4^q \mathbb{E}[\mathbb{E}_\sigma[Z_r^q]] \]

which \( Z'_r = \sup_f \sum_{(i,j) \in E} \sigma_i \sigma_j f(X_i, X_j) \). Note that \( \mathbb{E}_\sigma \) now denotes expectation taken with respect to both the \( \sigma \) and the \( \sigma' \). For simplicity, we denote by \( A = \mathbb{E}[Z'_r] \). In order to apply Corollary 3 of (Boucheron et al. 2005), define, for \( k = 1, \ldots, n \), the random variables
\[ A_k = \mathbb{E}[\sup_f \sum_{(i,j) \in E, i,j \neq k} w_{i,j} \sigma_i \sigma'_j f(X_i, X_j)]. \]

It is easy to see that \( A_k \leq A \).

On the other hand, defining
\[ R_k = \sup_f \sum_{i \in (i,k) \in E} w_{i,k} \sigma_i f(X_i, X_k), \]
\[ M = \max_k R_k \]

and denoting by \( f^* \) the function achieving the maximum in the definition of \( Z \), we clearly have
\[ A - A_k \leq 2 \mathbb{E}_\sigma[M] \]

and
\[ \sum_{k=1}^n (A - A_k) \leq 2A. \]

Therefore,
\[ \sum_{k=1}^n (A - A_k)^2 \leq 4A \mathbb{E}_\sigma[M]. \]

Then by Corollary 3 of (Boucheron et al. 2005), we obtain
\[ \mathbb{E}[\mathbb{E}_\sigma[Z'_r]^q] \leq 2^{q-1} (2^q \mathbb{E}[Z'_r]^q + 5^q \mathbb{E}[\mathbb{E}_\sigma[Z'_r]^q]) \]

(46)

To bound \( \mathbb{E}[\mathbb{E}[Z'_r]^q] \), observe that \( \mathbb{E}_\sigma[M] \) is a conditional Rademacher average, for which Theorem 13 of (Boucheron et al. 2005) could be applied. Since \( \max_k \sup_f \sum_{i \in (i,k) \in E} w_{i,k} f(X_i, X_k) \leq ||w||_\infty F', \) we have
\[ \mathbb{E}[\mathbb{E}_\sigma[M]^q] \leq 2^{q-1} (2^q \mathbb{E}[M]^q + 5^q ||w||_\infty^q F'^q). \]

(47)

By undecoupling, we have \( \mathbb{E}[Z'_r] = \mathbb{E}[\mathbb{E}_\sigma[Z'_r]] \leq 4 \mathbb{E}[\mathbb{E}_\sigma[Z_r]] = 4 \mathbb{E}[Z_r] \). Collecting all terms, we have
\[ \mathbb{E}[\mathbb{E}_\sigma[Z_r]^q] \leq 64^q \mathbb{E}[Z_r]^q + 16^q \mathbb{E}[M]^q \]
\[ + 400^q ||w||_\infty^q F'^q q^2. \]

**Second term:** \( \mathbb{E}[\mathbb{E}_\sigma[U_r]^q] \).

By the Cauchy-Schwartz inequality we can observe that
\[ \sup_{f, i} \sum_{\alpha : ||\alpha||_2 \leq 1} \sup_{j \in (i,k) \in E} \sum_{i,j} w_{i,j} \alpha_i \alpha'_j f(X_i, X_j) \leq ||w||_\max F. \]

Then similar to the bound of \( \mathbb{E}[\mathbb{E}_\sigma[M]^q] \), we have
\[ \mathbb{E}[\mathbb{E}_\sigma[U_r]^q] \leq 2^{q-1} (2^q \mathbb{E}[U_r]^q + 5^q ||w||_\max^q F'^q). \]

**Third term:** \( \mathbb{E}[B^q] \). By the Cauchy-Schwartz inequality, we have \( B \leq \sqrt{\sum_{(i,j) \in E} w_{i,j}^2 F} = ||w||_2 F \) so
\[ \mathbb{E}[B^q] \leq ||w||_2^q F^q. \]

Now it remains to simply put the pieces together to obtain
\[ \mathbb{E}[\sup_f \sum_{(i,j) \in E} w_{i,j} f(X_i, X_j)] \]
\[ \leq C (\mathbb{E}[Z_r]^q + q^2 \mathbb{E}[U_r]^q + q^2 \mathbb{E}[M]^q \]
\[ + ||w||_\max^q F^q q^2 + ||w||_\infty^q F'^q q^{3/2} \]
\[ + F^q ||w||_2^{2q}. \]
for an appropriate constant $C$. In order to derive the exponential inequality, we use Markov inequality $P[X \geq t] \leq t^{-\alpha} E[Z']$ and choose
\[
q = C \min \left( \left( \frac{t}{E[U' \sigma]} \right)^2, \frac{t}{E[M] F} \right) \frac{t}{\|w\|_\infty^2} \left( \frac{t}{\|w\|_\text{max} F} \right)^{2/3}, \sqrt{\frac{t}{\|w\|_\infty F}}
\]
for an appropriate constant $C$.

\textbf{Proof of Corollary 1.} From Theorem 3, it is easy to know
\[
\kappa = C(E[Z] + \max \{E[U] \sqrt{\log(1/\delta)}, E[M] \log(1/\delta), \log(1/\delta)\|w\|_2, (\log(1/\delta))^{3/2}\|w\|_\text{max}, (\log(1/\delta))^2\|w\|_\infty \}).
\]

An important character of these Rademacher processes in (48) is that they satisfy the Khintchine inequality (50). For simplicity, we denote the weighted Rademacher processes by $Z_\sigma^*$
\[
\{z_\sigma(f) = \sum_{(i,j) \in E} w_{ij} \sigma_i \sigma_j f(X_i, X_j), f \in F \}.
\]

where $E_\leq = \{(i, j) : (i, j) \in E, i < j \}$. Let $z_\sigma^* = z_\sigma^*/\|w\|_2$, following Theorem 8, we can easily have that $z_\sigma^*$ satisfies (50) with degree 2. Thus from Theorem 9, we have
\[
E_\sigma[sup f, g] z_\sigma^*(f) - z_\sigma^*(g) \leq K \int_0^D \log N(F, \mathbb{L}_d, \epsilon) d\epsilon.
\]

Recall that $sup F = F$, we have $D = 2F$. Since this metric function $d$ is intractable, we need to convert it to the $\mathbb{L}_\infty$ metric. For all $f, s$, we have $L_2(z_\sigma^*(f), z_\sigma^*(g)) \leq L_\infty(f, g)$. The fact that if $\forall f, s \in F, \mathbb{L}(f, s) \leq F(f, s)$ then $N(F, L, \epsilon) \leq N(F, L', \epsilon)$ combined with (49) will give
\[
E[Z] = 2E[f, g] z_\sigma(f) \|w\|_2 \int_0^{2F} \log N(F, \mathbb{L}_2, \epsilon) d\epsilon \\
\leq 2K \|w\|_2 \int_0^{2F} \log N(F, \mathbb{L}_2, \epsilon) d\epsilon \\
\leq 2K \|w\|_2 \int_0^{2F} \log N_\infty(F, \epsilon) d\epsilon
\]

where $K$ is a universal constant. Similarly, we can also bound $E[M]$ by $K\|w\|_\text{max}$
\[
\int_0^{2F} \log N_\infty(F, \epsilon) d\epsilon.
\]

for $U_\sigma$, we want to be the (random) vector that maximizes $U_\sigma$ and define
\[
\{u_\sigma(f) = \sum_{i=1}^n \sigma_i \sum_{(i,j) \in E} w_{ij} \sigma_i^* f(X_i, X_j), f \in F \}.
\]

Clearly, $u_\sigma$ satisfies the Khintchine inequality with degree 1. Also, we need to convert its metric distance and $L_2(u_\sigma^*(f), u_\sigma(g)) \leq L_\infty(f, g)$. Thus, $E[U_\sigma] \leq K\|w\|_2 \int_0^{2F} \log N_\infty(F, \epsilon) d\epsilon$.

Plugging all these part into (48) will complete the corollary.

\textbf{E Metric Entropy Inequality}

The following theorems are more or less classical and well known. We present them here for the sake of completeness.

\textbf{Theorem 8 (Khintchine inequality for Rademacher chaos, De la Pen and Giné (2012, Theorem 3.2.1)).} Let $F$ be a normed vector space and let $\{\sigma_i\}_{i=1}^\infty$ be a Rademacher sequence. Denote by
\[
X = x + \sum_{i=1}^n x_i \sigma_i + \sum_{i_1 < i_2 \leq n} x_{i_1, i_2} \sigma_{i_1} \sigma_{i_2} + \ldots + \sum_{i_1 < \cdots < i_d \leq n} x_{i_1, \ldots, i_d} \sigma_{i_1} \ldots \sigma_{i_d}
\]
the Rademacher chaos of order $d$. Let $1 < p \leq q < \infty$ and let
\[
\gamma = \left(\frac{p-1}{q-1}\right)^{1/2}.
\]

Then, for all $d \geq 1$,
\[
(\mathbb{E}|x + \sum_{i=1}^n \gamma x_i \sigma_i + \sum_{i_1 < i_2 \leq n} \gamma^2 x_{i_1, i_2} \sigma_{i_1} \sigma_{i_2} + \ldots + \sum_{i_1 < \cdots < i_d \leq n} \gamma^d x_{i_1, \ldots, i_d} \sigma_{i_1} \ldots \sigma_{i_d}|)^{1/p} \leq (\mathbb{E}|x + \sum_{i=1}^n x_i \sigma_i + \sum_{i_1 < i_2 \leq n} x_{i_1, i_2} \sigma_{i_1} \sigma_{i_2} + \ldots + \sum_{i_1 < \cdots < i_d \leq n} x_{i_1, \ldots, i_d} \sigma_{i_1} \ldots \sigma_{i_d}|)^{1/p}. 
\]

\textbf{Theorem 9 (metric entropy inequality, Arcones and Giné (1993, Proposition 2.6)).} If a process $\{Y_f : f \in F\}$ satisfies
\[
(\mathbb{E}[|Y_f - Y_g|^p])^{1/p} \leq (\frac{p-1}{q-1})^{m/2} (\mathbb{E}[|Y_f - Y_g|^q])^{1/q},
\]
for $1 < q < p < \infty$ and some $m \geq 1$, and if
\[
d(f, g) = (\mathbb{E}[|Y_f - Y_g|^2])^{1/2},
\]
there is a constant $K < \infty$ such that
\[
E[sup f, g] |Y_f - Y_g| \leq K \int_0^D (\log N(F, d, \epsilon))^{m/2} d\epsilon.
\]

where $D$ is the $d$-diameter of $F$. 

\textbf{References}

We refer the reader to the original papers for details.