On the ERM Principle with Networked Data

AAAI Press
Association for the Advancement of Artificial Intelligence
2275 East Bayshore Road, Suite 160
Palo Alto, California 94303

Abstract

Networked data, in which every training example involves two objects and multiple examples may share some common objects, is used in many machine learning tasks, e.g., learning to rank and link prediction. A challenge 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 on this problem are based on the classical ERM principle that always weights every example equally, but this strategy leads to unsatisfactory bounds. We consider general weighted ERM and provide new universal risk bounds for this problem, based on which the efficient methods to find good weights are also discussed.

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 do 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. Similar examples of networked data such as traffic networks (Min and Wynter 2011), chemical interaction networks (Letovsky and Kasif 2003), citation networks (McGovern et al. 2003) abound throughout the machine learning world.

Admittedly, many efforts have been made to design practical algorithms for learning from networked data, e.g., (Taskar et al. 2004, Al Hasan et al. 2006, Macskassy and Provost 2007). 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 (*CTND*)* 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 $X \in \mathcal{X}$ 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 graphs, and the bounds are independent of the distribution of the data. These bounds are of 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 hence G is usually incomplete, thus techniques based on complete U-processes derived from (Papa, Bellet, and Clémençon 2016) cannot be applied and the risk bounds of 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 order $O(\sqrt{\chi^*(D_G)/m})$ 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, Ramon, and Guo (2014) propose *weighted* ERM which adds weights to training examples according to the network, and show that the risk bound for weighted ERM can be of order $O(1/\sqrt{\nu^*(G)})$ where $\nu^*(G)$ 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 $\nu^*_G \geq m/\chi^*(D_G)$. However, Wang, Ramon, and Guo (2014) (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 (MT-noise condition)* (Mammen and Tsybakov 1998), which is commonly assumed in several typical learning problems with networked data, e.g., ranking (Clémençon, Lugosi, and Vayatis 2008) and graph reconstruction (Papa, Bellet, and

Clémençon 2016), can be used to reasonably bound the dependencies of two examples that share a common vertex and then leads to *tighter* risk bounds. Based on these new bounds, we can efficiently find better weighting schemes.

2 Preliminaries

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

2.1 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 $P_{\mathcal{X}}$. 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 $P_{\mathcal{Y}|\mathcal{X}^2}$ such that for every pair $(i, j) \in E$, the conditional probability

$$P\left[Y_{i,j} = y \mid x_1, \ldots, x_n\right] = P_{\mathcal{Y}|\mathcal{X}^2}\left[y, x_i, x_j\right].$$

Example 1 (pairwise ranking)**.** *(Liu and others 2009) categorize ranking problems into three groups by their input representations and loss functions. One of these categories is* pairwise ranking *that learns a binary classifier telling which document is better in a given pair of documents. A document can be described by a feature vector from the $\mathcal{X}$ describing title, volume, ... The target value (rank) between two documents, that only depends on features of these two documents, is* 1 *if the first document is considered better than the second, and* 0 *otherwise.*

The training set $S := \{(X_i, X_j, Y_{i,j})\}_{(i,j)\in E}$ is *dependent* copies of a generic random vector $(X_1, X_2, Y_{1,2})$ whose distribution $P = P_{\mathcal{X}} \otimes P_{\mathcal{X}} \otimes P_{\mathcal{Y}|\mathcal{X}^2}$ is fully determined by the pair $(P_{\mathcal{X}}, P_{\mathcal{Y}|\mathcal{X}^2})$. Let $\mathcal{R}$ be the set of all measurable functions from $\mathcal{X}^2$ to $\mathcal{Y}$ and for all $r \in \mathcal{R}$, the *loss function* $\ell(r, (x_1, x_2, y_{1,2})) = \mathbb{1}_{y_{1,2} \neq r(x_1, x_2)}$.

Given a graph G with training examples S and a *hypothesis set* $R \subseteq \mathcal{R}$, the CTND problem is to find a function $r \in R$, with risk

$$L(r) = \mathbb{E}[\ell(r, (X_1, X_2, Y_{1,2}))] \tag{1}$$

that achieves a comparable performance to the *Bayes rule* $r^* = \arg\inf_{r\in\mathcal{R}} L(r) = \mathbb{1}_{\eta(x_1,x_2)\geq 1/2}$, whose risk is denoted by L^*, where $\eta(x_1, x_2) = P_{\mathcal{Y}|\mathcal{X}^2}[1, x_1, x_2]$ is the *regression function*.

The main purpose of this paper is to devise a principle to select a classifier $\hat{r}$ from the hypothesis set R and establish bounds for its *excess risk* $L(\hat{r}) - L^*$.

Definition 1 ("low-noise" condition)**.** *Let us consider a learning problem, in which the hypothesis set is $\mathcal{F}$ and the Bayes rule is f^*. With slightly abusing the notation, this problem satisfies the "low-noise" condition if $\forall f \in \mathcal{F}, L(f) - L^* \geq C^\theta(\mathbb{E}[|f - f^*|])^\theta$ where C is a positive constant.*

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).

Definition 2 (symmetry)**.** *A learning problem is symmetric if for every $x_i, x_j \in \mathcal{X}$, $y_{i,j} \in \mathcal{Y}$ and $r \in R$, $\ell(r, (x_i, x_j, y_{i,j})) = \ell(r, (x_j, x_i, y_{j,i}))$.*

Many typical learning problems are symmetric. For example, pairwise ranking problem with symmetric functions r in the sense that $r(X_1, X_2) = 1 - r(X_2, X_1)$ satisfies the symmetric condition.

2.2 Weighted ERM

ERM aims to find the function from a hypothesis set that minimizes the empirical estimator of (1) on the training examples $S = \{(X_i, X_j, Y_{i,j})\}_{(i,j)\in E}$:

$$L_m(r) = \frac{1}{m} \sum_{(i,j)\in E} \ell(r, (X_i, X_j, Y_{i,j})). \tag{2}$$

where m is the number of training examples. In this paper, we consider its weighted version, in which we put weights on the examples and select the minimizer $r_{\mathbf{w}}$ of the weighted empirical risk

$$L_{\mathbf{w}}(r) = \frac{1}{\|\mathbf{w}\|_1} \sum_{(i,,j)\in E} w_{i,j}\ell(r, (X_i, X_j, Y_{i,j})) \tag{3}$$

where $\mathbf{w}$ is a *fractional matching* of G and $\|\mathbf{w}\|_1 > 0$.

Definition 3 (fractional matching)**.** *Given a graph $G = (V, E)$, a* fractional matching $\mathbf{w}$ *is a non-negative vector $(w_{i,j})_{(i,j)\in E}$ that for every vertex $i \in V, \sum_{j:(i,j)\in E} w_{i,j} \leq 1$.*

3 Intuitions and Examples

We now have a look at previous works that are closely related to our work, as shown in Table 1, and present the merits of our method. Biau and Bleakley (2006), Clémençon, Lugosi, and Vayatis (2008) and Papa, Bellet, and Clémençon (2016) deal with the case when the graph is complete, i.e., the target value of every pair of vertices is known. In this case, Clémençon, Lugosi, and Vayatis (2008) formulate the "low-noise" condition for the ranking problem and demonstrate that this condition can lead to tighter risk bounds by the moment inequality of U-processes. Papa, Bellet, and Clémençon (2016) further consider the graph reconstruction problem introduced by Biau and Bleakley (2006) and show this problem always satisfies the "low-noise" condition.

If the graph is incomplete, one can use either Janson's decomposition (Janson 2004, Usunier, Amini, and Gallinari 2006, Ralaivola, Szafranski, and Stempfel 2009, Ralaivola and Amini 2015) or the fractional matching approach by Wang, Ramon, and Guo (2014) to derive risk bounds. The main differences between these two approaches are:

Table 1: Summary of methods for CTND.

Principles	Graph type	With "low-noise" condition	Without "low-noise" condition
Unweighted ERM (equally weighted)	**Complete graphs**	Clémençon, Lugosi, and Vayatis (Ann. Stat. 2008), Papa, Bellet, and Clémençon (NIPS 2016)	Biau and Bleakley (Statistics and Decisions 2006)
	General graphs	Ralaivola and Amini (ICML 2015)	Usunier, Amini, and Gallinari (NIPS 2006), Ralaivola, Szafranski, and Stempfel (AISTATS 2009)
Weighted ERM	**General graphs**	*This paper*	(Wang, Ramon, and Guo 2014)

- Wang, Ramon, and Guo (2014) considers the data graph G while Janson's decomposition uses only the line graph D_G.
- The fractional matching approach considers general weighted ERM while Janson (2004), Usunier, Amini, and Gallinari (2006), Ralaivola, Szafranski, and Stempfel (2009) and Ralaivola and Amini (2015) only prove bounds for unweighted ERM.

Though Wang, Ramon, and Guo (2014) show improved risk bounds, as far as we know, there is no known tight risk bounds on incomplete graphs for tasks such as pairwise ranking and graph reconstruction that satisfy the "low-noise" condition, under which the weighting scheme proposed by Wang, Ramon, and Guo (2014) may be not good either (see Section 5.2).

Line graphs Compared to Janson's decomposition, 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 neglect makes it unable to improve bounds, no matter whether considering weighted ERM (see Section A.1 in the appendix). In Section 5.3, we show that our bounds are tighter than that of line graph based methods.

Asymptotic risk As mentioned by Wang, Ramon, and Guo (2014), 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 risk bounds that do not converge to 0 (see the example in Section 5.3).

Uniform bounds Although Ralaivola and Amini (2015) prove an entropy-base concentration inequality for networked data using Janson's decomposition, the condition it required is 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 Clémençon 2016)) to establish uniform bounds, in absence of any restrictive condition imposed on the data distribution.

4 Main Result

4.1 Covering Numbers

The excess risk $L(r_{\mathbf{w}}) - L^*$ depends on the hypothesis set R whose complexity is measured by *covering number* (Cucker and Zhou 2007) in this paper. A similar result using VC-dimension (Vapnik and Chervonenkis 1971) can be obtained as well.

Definition 4 (covering numbers)**.** *$\mathcal{C} \subset \mathbb{R}^m$ is a d_p cover of $\mathcal{F}$ on $x_1, \ldots, x_m$ at scale $\epsilon > 0$ if for all $f \in \mathcal{F}$, there exists $\mathbf{c}_f \in \mathcal{C}$ such that $\|(f(x_1), \ldots, f(x_m)) - \mathbf{c}_f\|_p \leq \epsilon$. The empirical covering number $N_p(\mathcal{F}, \epsilon; x_1, \ldots, x_m)$ is the minimum cardinality of $\mathcal{C}$ such that $\mathcal{C}$ is a d_p cover of $\mathcal{F}$ on $x_1, \ldots, x_m$ at scale ϵ. We define the covering number $N_p(\mathcal{F}, \epsilon, m) = \sup_{x_1,\ldots,x_m} N_p(\mathcal{F}, \epsilon; x_1, \ldots, x_m)$. We simply denote $N_p(\mathcal{F}, \epsilon, m)$ by $N_p(\mathcal{F}, \epsilon)$, if the context is clear.*

In this paper, we focus on the $\mathbb{L}_\infty$ covering number $N_\infty(\mathcal{F}, \epsilon)$ and assume that it satisfies the following assumption. 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 $\log N_\infty(\mathcal{F}, \epsilon)$ through 0.

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

By Sauer's lemma (Sauer 1972), one can easily prove a binary function class with VC-dimension V satisfies Assumption 1 with $K = V \log(em/V)$ and $\beta = 0$. Besides, Dudley (1974), Korostelev and Tsybakov (1993) and Mammen and Tsybakov (1995) have given various examples of classes $\mathcal{F}$ satisfying Assumption 1. We also refer to (Mammen and Tsybakov 1998, p. 1813) for concrete examples of classes of the hypothesis sets with smooth boundaries satisfying Assumption 1.

4.2 Risk Bounds

Now we are ready to state the main result of the paper.

Theorem 1 (risk bounds)**.** *Consider a minimizer $r_{\mathbf{w}}$ of the weighted empirical risk $L_{\mathbf{w}}$ over a class R that satisfies Assumption 1. There exists a universal constant C such that for all $\delta \in (0,1]$, with probability at least $1-\delta$, the excess risk of $r_{\mathbf{w}}$ satisfies*

$$L(r_{\mathbf{w}}) - L^* \leq 2(\inf_{r\in R} L(r) - L^*) + \frac{KC\log(1/\delta)}{(1-\beta)^{2/(\beta+1)}\|\mathbf{w}\|_1} \left(\|\mathbf{w}\|_1^{\beta/(1+\beta)} + \max\Big(\|\mathbf{w}\|_2, \|\mathbf{w}\|_{\max}(\log(1/\delta))^{1/2}, \|\mathbf{w}\|_\infty(\log(1/\delta)) \Big) \right) \tag{4}$$

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

From Theorem 1, for all $\delta \in \left(\exp\Big(- \min(\|\mathbf{w}\|_2/\|\mathbf{w}\|_\infty, \|\mathbf{w}\|_2^2/\|\mathbf{w}\|_{\max}^2) \Big), 1 \right]$, the risk bounds are of the order $O\left((1/\|\mathbf{w}\|_1)^{1/(1+\beta)} + \|\mathbf{w}\|_2/\|\mathbf{w}\|_1\right)$. In this case, our bounds are tighter than $O(1/\sqrt{\|\mathbf{w}\|_1})$ (recall that $\mathbf{w}$ must be a fractional matching) as $\|\mathbf{w}\|_2/\|\mathbf{w}\|_1 \leq 1/\sqrt{\|\mathbf{w}\|_1}$. If G is complete and every example is unweighted, the bounds of the order $O((1/n)^{1/(1+\beta)})$ can achieve the same results in (Papa, Bellet, and Clémençon 2016) (with the same condition on δ but different complexity measurements of the hypothesis set).

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 price of increasing the constant C. Wang, Ramon, and Guo (2014) obtain risk bounds that has a factor* 1 *in front of the approximation error part, but in their result the bound is $O(1/\sqrt{\|\mathbf{w}\|_1})$. 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 (the detailed proofs can be found in Section D in the appendix). We first define

$$q_r(x_1, x_2, y_{1,2}) := \ell(r, x_1, x_2, y_{1,2}) - \ell(r^*, x_1, x_2, y_{1,2})$$

for every $(x_1, x_2, y_{1,2}) \in \mathcal{X}\times\mathcal{X}\times\mathcal{Y}$ and let $\Lambda(r) = L(r) - L^* = \mathbb{E}[q_r(X_1, X_2, Y_{1,2})]$ be the excess risk with respect to the Bayes rule. Its empirical estimate by weighted ERM is

$$\begin{aligned}\Lambda_{\mathbf{w}}(r) &= L_{\mathbf{w}}(r) - L_{\mathbf{w}}(r^*) \\ &= \frac{1}{\|\mathbf{w}\|_1} \sum_{(i,j)\in E} w_{i,j} q_r(X_i, X_j, Y_{i,j}).\end{aligned}$$

By Hoeffding's decomposition (Hoeffding 1948), for all $r \in \mathcal{R}$, one can write

$$\Lambda_{\mathbf{w}}(r) = T_{\mathbf{w}}(r) + U_{\mathbf{w}}(r) + \widetilde{U}_{\mathbf{w}}(r), \tag{5}$$

where

$$T_{\mathbf{w}}(r) = \Lambda(r) + \frac{2}{\|\mathbf{w}\|_1} \sum_{i=1}^n \sum_{j:(i,j)\in E} w_{i,j} h_r(X_i)$$

is a weighted average of i.i.d. random variables with $h_r(X_i) = \mathbb{E}[q_r(X_i, X_j, Y_{i,j}) \mid X_i] - \Lambda(r)$,

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

is a weighted *degenerated* U-statisitc of order 2 with symmetric kernel $\hat{h}_r(X_i, X_j) = \mathbb{E}[q_r(X_i, X_j, Y_{i,j}) \mid X_i, X_j] - \Lambda(r) - h_r(X_i) - h_r(X_j)$ and

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

with a *degenerated* kernel $\tilde{h}_r(X_i, X_j, Y_{i,j}) = q_r(X_i, X_j, Y_{i,j}) - \mathbb{E}[q_r(X_i, X_j, Y_{i,j}) \mid X_i, X_j]$. In the following, we bound these three terms $T_{\mathbf{w}}, U_{\mathbf{w}}$ and $\widetilde{U}_{\mathbf{w}}$ respectively.

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

$$\sup_{r\in R} |U_{\mathbf{w}}(r)| \leq \frac{\max(K, \sqrt{K}) C_1}{1-\beta} \max\Big(\frac{\|\mathbf{w}\|_2 \log(1/\delta)}{\|\mathbf{w}\|_1}, \frac{\|\mathbf{w}\|_{\max}(\log(1/\delta))^{3/2}}{\|\mathbf{w}\|_1}, \frac{\|\mathbf{w}\|_\infty(\log(1/\delta))^2}{\|\mathbf{w}\|_1} \Big)$$

and

$$\sup_{r\in R} |\widetilde{U}_{\mathbf{w}}(r)| \leq \frac{\max(K, \sqrt{K}) C_2}{1-\beta} \left(\frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_1} + \max\Big(\frac{\|\mathbf{w}\|_{\max}(\log(1/\delta))^{3/2}}{\|\mathbf{w}\|_1}, \frac{\|\mathbf{w}\|_\infty(\log(1/\delta))^2}{\|\mathbf{w}\|_1} \Big) \right)$$

where $C_1, C_2 < +\infty$ is a universal constant.

To prove Lemma 1, we show that $U_{\mathbf{w}}(r)$ and $\widetilde{U}_{\mathbf{w}}(r)$ can be bounded by Rademacher chaos using classical symmetrization and randomization tricks combined with the decoupling method (see Section B in the Appendix). We handle these Rademacher chaos by generalizing the approach used in (Clémençon, Lugosi, and Vayatis 2008). Specifically, we utilize the moment inequalities from (Boucheron et al. 2005) to convert them to the sum of simpler processes, which can be bounded by the metric entropy inequality for Khinchine-type processes (see Dembo, Karlin, and Zeitouni 1994, Proposition 2.6) and Assumption 1.

Lemma 1 shows that the contribution of the degenerated parts $U_{\mathbf{w}}(r)$ and $\widetilde{U}_{\mathbf{w}}(r)$ to the excess risk can be bounded. This implies that minimizing $\Lambda_{\mathbf{w}}(r)$ is approximately equivalent to minimizing $T_{\mathbf{w}}(r)$ and thus $r_{\mathbf{w}}$ is a ρ-minimizer of

$T_{\mathbf{w}}(r)$ in the sense that $T_{\mathbf{w}}(r_{\mathbf{w}}) \leq \rho + \inf_{r \in R} T_{\mathbf{w}}(r)$. In order to analyze $T_{\mathbf{w}}(r)$, 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 A in the Appendix). Based on this result, tight bounds for the excess risk with respect to $T_{\mathbf{w}}(r)$ can be obtained if the variance of the excess risk is controlled by its expected value. By Lemma 2, $T_{\mathbf{w}}(r)$ fulfills this condition.

Lemma 2 (condition leads to "low-noise", Papa, Bellet, and Clémençon (2016, Lemma 2))**.** *If the learning problem* CTND *is symmetric, then*

$$Var\left[\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1]\right] \leq \Lambda(r) \tag{6}$$

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 ρ-minimizer of $T_{\mathbf{w}}(r)$ in the sense that $T_{\mathbf{w}}(r') \leq \rho + \inf_{r \in R} T_{\mathbf{w}}(r)$ and R satisfies Assumption 1, then there exists a universal constant C such that for all $\delta \in (0, 1]$, with probability at least $1 - \delta$, the risk of r' satisfies*

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

Plugging the result of Lemma 1 into Lemma 3, we can prove Theorem 1. 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 $\|\mathbf{w}\|_1$ (under the constraint that this weight vector is a fractional matching), which resembles the result of (Wang, Ramon, and Guo 2014) (but they only need to maximize $\|\mathbf{w}\|_1$ and this is why they end in the $O(1/\sqrt{\nu^*(G)})$ bound), while making $\|\mathbf{w}\|_2, \|\mathbf{w}\|_{\max}, \|\mathbf{w}\|_\infty$ as small as possible, which appears to suggest putting nearly average weights on examples and vertices respectively. These two objectives, maximizing $\|\mathbf{w}\|_1$ and minimizing $\|\mathbf{w}\|_2, \|\mathbf{w}\|_{\max}, \|\mathbf{w}\|_\infty$, seem to contradict each other. In the next section, we discuss how to solve this problem.

5 Weighting Schemes

In the previous section, we have established risk bounds for weighted ERM. In this section, we first formulate the optimization problem that minimizes the risk bounds in Theorem 1. We show that, according to our bounds, equal weighting is indeed the optimal weighting scheme for complete graphs. We then discuss the performance of this equal weighting scheme when the graph is incomplete. Finally, we provide some methods that optimize the risk bounds progressively to find the proper weight for general cases. Using these methods, we are able to improve existing weighting schemes.

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:

$$\begin{aligned} \min_{\mathbf{w}} \quad & \frac{1}{\|\mathbf{w}\|_1} \Bigg(\|\mathbf{w}\|_1^{\beta/(1+\beta)} + \max\Big(\|\mathbf{w}\|_2, \\ & \qquad \|\mathbf{w}\|_{\max} (\log(1/\delta))^{1/2}, \|\mathbf{w}\|_\infty (\log(1/\delta)) \Big) \Bigg) \\ \text{s.t.} \quad & \forall (i,j) \in E, w_{i,j} \geq 0 \quad \text{and} \quad \forall i, \sum_{j:(i,j)\in E} w_{i,j} \leq 1 \end{aligned} \tag{7}$$

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

$$\begin{aligned} \min_{\mathbf{p}} \quad & \Big(\max_{i=1,\ldots,n} \sum_{j:(i,j)\in E} p_{i,j} \Big)^{1/(1+\beta)} + \max\Big(\|\mathbf{p}\|_2, \\ & \qquad \|\mathbf{p}\|_{\max} (\log(1/\delta))^{1/2}, \|\mathbf{p}\|_\infty (\log(1/\delta)) \Big) \\ \text{s.t.} \quad & \forall (i,j) \in E, p_{i,j} \geq 0 \quad \text{and} \quad \sum_{(i,j)\in E} p_{i,j} = 1 \end{aligned} \tag{8}$$

5.2 Complete Graphs

When graph G is complete, it is easy to see that weighting all examples equally gives the best risk bound, as all the terms $\max_{i=1,\ldots,n} \sum_{j:(i,j)\in E} p_{i,j}$, $\|\mathbf{p}\|_2$, $\|\mathbf{p}\|_{\max}$ and $\|\mathbf{p}\|_\infty$ achieve minimum. Compared to the results in (Wang, Ramon, and Guo 2014), our theory puts additional constrains on $\|\mathbf{p}\|_2$, $\|\mathbf{p}\|_{\max}$ and $\|\mathbf{p}\|_\infty$ 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 (Hoeffding 1948), could produce a small variance.

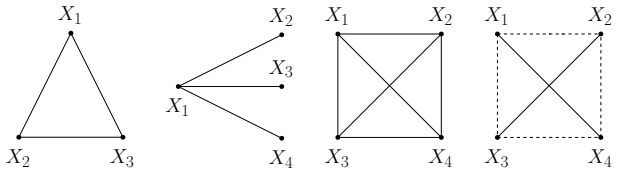

Figure 1: (a) and (b) are two different data graphs while both of them correspond to the same line graph (a triangle). (c) and (d) are two weighting schemes for a complete graph formed by points X_1, X_2, X_3, X_4. Solid line means its weight $p > 0$ while dash line means $p = 0$. (c): Weight every example equally. (d): Only the two examples in a independent subset get equally non-zero weights and other weights are 0 (dashed line). Note that $\max_{i=1,\ldots,n} \sum_{j:(i,j)\in E} p_{i,j}$ of these two weighting schemes are the same, but (c) has the tighter risk bounds, as $\|\mathbf{p}\|_2$, $\|\mathbf{p}\|_{\max}$ and $\|\mathbf{p}\|_\infty$ of (c) are smaller than that of (d) respectively.

5.3 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

$$O((\frac{\Delta(G)}{m})^{1/(1+\beta)} + \frac{1}{\sqrt{m}}). \tag{9}$$

if $\delta \in (\exp(-m/\Delta(G)), 1]$. In some cases, such as bounded degree graphs and complete graphs, this scheme could provide tight 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, we show that these bounds are always better than the Janson's decomposition, by which the risk bounds are of the order $O(\sqrt{\chi^*(D_G)/m})$.

However, as argued in Section 3, one can construct examples to illustrate that if $\Delta(G)$ is large (e.g., it is linear to m) and if we use equal weighting strategy, the risk bounds (9) 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 order $O(1)$ that is meaningless. A much better weighting scheme of this case is to weight the examples of disjoint edges with 1 while weight the examples of adjacent edges with $2/m$, which provides risk bounds of the order $O\left((1/m)^{1/(1+\beta)} + \sqrt{1/m}\right)$*[1].

5.4 Optimization methods

We rewrite the program (8) by introducing two auxiliary variables a and b:

$$\begin{aligned}
\min_{a,b,\mathbf{p}} \quad & a^{1/(1+\beta)} + b \\
\text{s.t.} \quad & \forall (i,j) \in E, p_{i,j} \geq 0 \\
& \forall (i,j) \in E, p_{i,j} \log(1/\delta) - b \leq 0 \\
& \forall i, \sum_{j:(i,j)\in E} p_{i,j} - a \leq 0 \\
& \forall i, \left(\sum_{j:(i,j)\in E} p_{i,j}^2 \log(1/\delta) \right)^{1/2} - b \leq 0 \\
& \|\mathbf{p}\|_2 - b \leq 0 \quad \text{and} \quad \sum_{(i,j)\in E} p_{i,j} = 1
\end{aligned} \tag{10}$$

Note that the constraints are all convex. If the hypothesis set is a VC class or even finite, i.e., $\beta = 0$, then the objective function becomes linear and then (10) is a convex optimization problem that can be solved by standard convex optimization methods (see e.g., (Boyd and Vandenberghe 2004)).

If $\beta > 0$, the objective function is not convex any more. In fact, the program (10) becomes a concave problem that can be optimized globally by some complex algorithms (Benson 1995, Hoffman 1981) that often need tremendous computation. Instead, one may not need to find the global optimum and only approximate it using some efficient methods.

Coordinate Descent Note that if a (or b) is fixed, then the objective function of (10) is monotonic (thus convex because we can remove the exponent $1/(1+\beta)$ of a), we can solve (10) by coordinate descent which optimizes a and b successively. That is, starting with a feasible solution, at each step, we solve the primal optimization problem (10) by fixing a or b at its value from the current iteration and minimizing the objective with respect to the remaining components. Since each iteration is a convex optimization problem, we can solve it by some standard methods (Boyd and Vandenberghe 2004). Clearly, the new solution $(a^{(t+1)})^{1/(1+\beta)} + b^{(t)}$ (respectively $(a^{(t+1)})^{1/(1+\beta)} + b^{(t+1)}$) is better than the old one $(a^{(t)})^{1/(1+\beta)} + b^{(t)}$ (respectively $(a^{(t+1)})^{1/(1+\beta)} + b^{(t)}$).

Concave-Convex Procedure Since the objective function of (10) can be rewritten to $b - (-a^{1/(1+\beta)})$ that is difference of convex function (d.c.), one can also solve (10) by the elaborate optimization methods for d.c. program. One of these methods is Concave-Convex Procedure (Yuille 2001) which has been wildly used in sparse support vector machine (SVM) and transductive SVM. Using this method, the program (10) can be solved by optimizing a sequence of convex program with the same constrains of (10), the iterative objective functions $(a^{(t+1)}, b^{(t+1)}) \in \arg\min_{a,b,\mathbf{p}} b + \frac{a}{1+\beta}(a^{(t)})^{-\beta/(1+\beta)}$ and the feasible initial solution. By Lanckriet and Sriperumbudur (2009), the concavity of $a^{1/(1+\beta)}$ guarantees that $(a^{(t+1)})^{1/(1+\beta)} + b^{(t+1)} \leq (a^{(t)})^{1/(1+\beta)} + b^{(t)}$.

6 Conclusion

In this paper, we consider weighted ERM of the symmetric CTND problem and establish tight universal risk bounds under the "low-noise" condition. There 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 better generalization by putting proper weight on every networked example. 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. We also propose efficient optimization algorithms to improve the existing weighting schemes in general case.

References

Al Hasan, M.; Chaoji, V.; Salem, S.; and Zaki, M. 2006. Link prediction using supervised learning. In *SDM06: workshop on link analysis, counter-terrorism and security*.

Benson, H. P. 1995. Concave minimization: theory, applications and algorithms. In *Handbook of global optimization*. Springer. 43–148.

[1] This example is only used to show the necessity of weighting, and the order cannot be improved by any approach.

Biau, G., and Bleakley, K. 2006. Statistical inference on graphs. *Stat. Decis.* 24(2):209–232.

Boucheron, S.; Bousquet, O.; Lugosi, G.; and Massart, P. 2005. Moment inequalities for functions of independent random variables. *Ann. Probab.* 33(2):514–560.

Bousquet, O. 2002. A Bennett concentration inequality and its application to suprema of empirical processes. *C. R. Acad. Sci. Paris, Ser. I* 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. *The Annals of Statistics* 844–874.

Cucker, F., and Zhou, D. X. 2007. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press.

de la Peña, V., and Giné, E. 1999. *Decoupling: From Dependence to Independence*. Springer Science & Business Media.

Dembo, A.; Karlin, S.; and Zeitouni, O. 1994. Limit Theorems for U-processes. *Ann. Probab.* 22(4):2022–2039.

Dudley, R. M. 1974. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory* 10(3):227–236.

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.

Lanckriet, G. R., and Sriperumbudur, B. K. 2009. On the convergence of the concave-convex procedure. In *Advances in neural information processing systems*, 1759–1767.

Letovsky, S., and Kasif, S. 2003. Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics* 19(suppl 1):i197–i204.

Liu, T.-Y., et al. 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. *Ann. Stat.* 34(5):2326–2366.

McGovern, A.; Friedland, L.; Hay, M.; Gallagher, B.; Fast, A.; Neville, J.; and Jensen, D. 2003. Exploiting relational structure to understand publication patterns in high-energy physics. *ACM SIGKDD Explorations Newsletter* 5(2):165–172.

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. 2015. Entropy-based concentration inequalities for dependent variables. In *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, 2436–2444.

Ralaivola, L.; Szafranski, M.; and Stempfel, G. 2009. Chromatic pac-bayes bounds for non-iid data. In *AISTATS*, 416–423.

Rejchel, W. 2012. On ranking and generalization bounds. *Journal of Machine Learning Research* 13(May):1373–1392.

Sauer, N. 1972. On the density of families of sets. *Journal of Combinatorial Theory, Series A* 13(1):145–147.

Taskar, B.; Wong, M.-F.; Abbeel, P.; and Koller, D. 2004. Link prediction in relational data. In *Advances in neural information processing systems*, 659–666.

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. *Advances in neural information processing systems* 18:1369.

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.; Ramon, J.; and Guo, Z.-C. 2014. Learning from networked examples. *arXiv preprint arXiv:1405.2600*.

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 the tight 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 $\widetilde{U}_{\mathbf{w}}(r)$. Then, some useful inequalities including the moment inequality are proved for weighted Rademacher chaos. Section D and E provide technical proofs for the main track and

the appendix. Section F introduces the Khinchine inequality and gives the metric entropy inequality for Rademacher chaos.

A Janson's decomposition

There are some literature that uses 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 5 (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 $e_i \cap e_j \neq \emptyset$ in G.*

This framework differs from our setting and detains less information from the data graph, as argued in Section 1.1. One can annalyze this framework by the fractional coloring and the Janson's decomposition.

Definition 6 (fractional coloring)**.** *Let $D_G = (D_V, D_E)$ be a graph. $\mathcal{C} = \{(\mathcal{C}_j, q_j)\}_{j\in(1,\dots,J)}$, for some positive integer J, with $\mathcal{C}_j \subseteq D_V$ and $q_j \in [0, 1]$ is an fractional coloring of D_G, if*

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

The weight $W(\mathcal{C})$ of $\mathcal{C}$ is given by $W(\mathcal{C}) = \sum_{j=1}^{J} q_j$ and the minimum weight $\chi^*(D_G) = \min_{\mathcal{C}} W(\mathcal{C})$ 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 fraction coloring of D_G and then analyze each set using the standard method for i.i.d. examples.

In particular, Theorem 2 (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 $\mathcal{Z}^m$ and $S_{\mathcal{C}_j}$ be the examples included in $\mathcal{C}_j$. Also, let k_j^i be the index of the i-th example of $\mathcal{C}_j$ in the training set S and denote by $\mathbf{w} = \{w_i\}_{i=1}^m$ the weights on S.

Theorem 2. *Using the notations defined above, let $\mathcal{C} = \{(\mathcal{C}_j, q_j)\}_{j=1}^J$ be a fractional coloring of D_G. Let $f : \mathcal{Z}^m \to \mathbb{R}$ such that*

- *there exist J functions $\mathcal{Z}^{|\mathcal{C}_j|} \to \mathbb{R}$ which satisfy $\forall S \in \mathcal{Z}^m$, $f(S) = \sum_{j=1}^J q_j f_j(S_{\mathcal{C}_j})$.*
- *there exist $c_i, \dots, c_m \in \mathbb{R}_+$ such that $\forall j, \forall S_{\mathcal{C}_j}, S^k_{\mathcal{C}_j}$ such that $S_{\mathcal{C}_j}$ and $S^k_{\mathcal{C}_j}$ differ only in the k-th example, $|f_j(S_{\mathcal{C}_j}) - f_j(S^k_{\mathcal{C}_j})| \leq c_{k_j^i}$.*

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 3. *Consider a minimizer $r_{\mathbf{w}}$ of the weighted empirical risk $L_{\mathbf{w}}$ over a class R. For all $\delta \in (0, 1]$, with probability at least $1 - \delta$, we have*

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

where

$$\mathfrak{R}^*_{\mathbf{w}}(R, S) = \frac{2}{N}\mathbb{E}_\sigma\left[\sum_{j=1}^J q_j \sup_{r\in R}\sum_{i\in\mathcal{C}_j} w_{k_j^i}\sigma_i r(z_{k_j^i})\right].$$

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

In this setting, although the weights $\mathbf{w}$ 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 $\sqrt{\chi^*(D_G)/m}$, as $\|\mathbf{w}\|_2/\|\mathbf{w}\|_1 \leq 1/\sqrt{m}$.

A.2 "Low-noise" condition

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

$$\frac{1}{m!}\sum_{\boldsymbol{\pi}\in\mathcal{G}_n}\frac{1}{\lfloor m/2\rfloor}\sum_{i=1}^{\lfloor m/2\rfloor}\ell(r, (X_{\pi(i)}, X_{\pi(i)+\lfloor m/2\rfloor}, Y_{\pi(i),\pi(i)+\lfloor m/2\rfloor})) \quad (11)$$

where the sum is taken over all permutations of $\mathcal{G}_m$, the symmetric group of order m, and $\lfloor u\rfloor$ denotes the integer part of any $u \in \mathbb{R}$. From (Biau and Bleakley 2006), the bounds for excess risk of (11) are of the order $O(1/\sqrt{n})$. Moreover, by the result in (Ralaivola and Amini 2015), tighter risk bounds may be obtained under the following "low-noise" condition (Tsybakov 2004, Massart and Nédélec 2006).

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

$$P\left[|\eta(X_1, X_2) - \frac{1}{2}| \leq \epsilon\right] \leq C\epsilon^{\theta/(1-\theta)}.$$

The risk bounds of the order $O(\log(n)/n)$ may be achieved if $\theta = 1$ (Massart and Nédélec 2006), which however is very restrictive. We can use the example in Papa, Bellet, and Clémençon (2016) to show this. Let N be a positive integer. For each vetex $i \in \{1, \dots, n\}$, we observe $X_i = (X_i^1, X_i^2)$ where X_i^1 and X_i^2 are two distinct elements drawn from $\{1, \dots, N\}$. This may, for instance correspond to the two preferred items of a user i among a list of N items. Consider now the case that two nodes are likely to be connected if they share common preference, e.g., $Y_{i,j} \sim Ber(\#(X_i \cap X_j)/2)$. One can easily check that $P[|\eta(X_1, X_2) - 1/2| = 0] > 0$, so tight risk bounds cannot be obtained for minimizers of (11).

B Tight bounds for weighted ERM on i.i.d. examples

In section 3, the excess risk is split into two types of processes: the weighted empirical process of i.i.d. examples and two degenerated processes. In this section, we prove general risk bounds for the weighted empirical process of i.i.d. examples. The main idea is similar to (Massart and Nédélec 2006) that tighter bounds for the excess risk can be obtained if the variance of the excess risk is controlled by its expected value.

B.1 Bennett concentration inequality

First, we prove a concentration inequality for the supremum of weighted empirical processes derived from (Bousquet 2002).

Theorem 4. *Assume the $(X_1, \ldots, X_i)$ are i.i.d. random variable according to P. Let $\mathcal{F}$ be a countable set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume that all functions f in $\mathcal{F}$ are P-measurable and square-integrable. If $\sup_{f\in\mathcal{F}}|f|\leq b$, we denote*

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

which $\{w_i\}_{i=1}^n$ are bounded weights such that $0 \leq w_i \leq 1$ for $i = 1, \ldots, n$ and $\|\mathbf{w}\|_1 > 0$. Let σ be a positive real number such that $\sigma^2 \geq \sup_{f\in\mathcal{F}} Var[f(X)]$ almost surely, then for all $x \geq 0$, we have

$$P\left[Z - \mathbb{E}[Z] \geq \sqrt{\frac{2(\frac{\|\mathbf{w}\|_2^2}{\|\mathbf{w}\|_1}\sigma^2 + 4b\mathbb{E}[Z])x}{\|\mathbf{w}\|_1}} + \frac{2bx}{3\|\mathbf{w}\|_1}\right] \leq e^{-x}. \tag{12}$$

It is a variant of Theorem 2.3 of (Bousquet 2002) by just applying Theorem 2.1 of (Bousquet 2002) with the weighted empirical process. Then, by Theorem 4, we can generalize the results of (Massart and Nédélec 2006) to weighted ERM. We start by describing the probabilistic framework adapts to our problem.

B.2 General upper bounds

Suppose that one observes independent variable $\xi_1, \ldots, \xi_n$ taking their values in some measurable space $\mathcal{Z}$ with commom distribution P. For every i, the variable $\xi_i = (X_i, Y_i)$ is a copy of a pair of random variables (X, Y) where X take its values in measurable space $\mathcal{X}$. Think of $\mathcal{R}$ as being the set of all measurable functions from $\mathcal{X}$ to $\{0, 1\}$. Then we consider some *loss* function

$$\gamma : \mathcal{R} \times \mathcal{Z} \to [0, 1]. \tag{13}$$

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

$$\bar{\ell}(r^*, r) = P[\gamma(r, \cdot) - \gamma(r^*, \cdot)], \forall r \in \mathcal{R} \tag{14}$$

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

$$\gamma_{\mathbf{w}}(r) = \frac{1}{\|\mathbf{w}\|_1} \sum_{i=1}^{n} w_i \gamma(r, \xi_i) \tag{15}$$

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

We introduce the *weighted centered empirical process* $\overline{\gamma_{\mathbf{w}}}$ defined by

$$\overline{\gamma_{\mathbf{w}}}(r) = \gamma_{\mathbf{w}}(r) - P[\gamma(r, \cdot)]. \tag{16}$$

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

$$\mathrm{Var}[\gamma(r, \cdot) - \gamma(r^*, \cdot)] \leq d^2(r, r^*), \forall r \in \mathcal{R}. \tag{17}$$

A tighter risk bound for weighted ERM is derived from Theorem 4 which combines two different moduli of uniform continuity: the stochastic modulus of uniform continuity of $\overline{\gamma_{\mathbf{w}}}$ 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 7. *We denote by D the class of nondecreasing and continuous functions ψ 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 $\xi \in \mathcal{Z}$, $\gamma(r_k, \xi)$ tends to $\gamma(r, \xi)$ as k tends to infinity.*

The upper bound for the relative expected loss of any empirical risk minimizer on some given model R will depend on the bias term $\bar{\ell}(r^*, R) = \inf_{r\in R} \bar{\ell}(r^*, r)$ and the fluctuations of the empirical process $\overline{\gamma_{\mathbf{w}}}$ on R. As a matter of fact, we shall consider some slightly more general estimators. Namely, given some nonnegative number ρ, we consider some ρ-empirical risk minimizer, that is, any estimator r taking its values in R such that $\gamma_{\mathbf{w}}(\hat{r}) \leq \rho + \inf_{r\in R} \gamma_{\mathbf{w}}(r)$.

Theorem 5 (risk bound for weighted ERM)**.** *Let γ be a loss function such r^* minimizes $P[\gamma(r, \cdot)]$ when r varies in $\mathcal{R}$. Let ϕ and ψ belong to the class of functions D defined above and let R be a subset of $\mathcal{R}$ satisfying the sparability Assumption 3. Assume that, on the one hand,*

$$d(r^*, r) \leq \frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2} \psi(\sqrt{\bar{\ell}(r^*, r)}), \forall r \in \mathcal{R}, \tag{18}$$

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

$$\sqrt{\|\mathbf{w}\|_1}\mathbb{E}\left[\sup_{r'\in R', \frac{\|\mathbf{w}\|_2}{\sqrt{\|\mathbf{w}\|_1}} d(r', r) \leq \sigma} [\overline{\gamma_{\mathbf{w}}}(r') - \overline{\gamma_{\mathbf{w}}}(r)\right] \leq \phi(\sigma) \tag{19}$$

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

$$\sqrt{\|\mathbf{w}\|_1}\epsilon_*^2 = \phi(\psi(\epsilon_*)). \tag{20}$$

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

$$P\left[\bar{\ell}(r^*, \hat{r}) > 2\rho + 2\bar{\ell}(r^*, R) + Ky\epsilon_*^2\right] \leq e^{-y}. \tag{21}$$

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 $P_{\mathbf{w}}(f) = \frac{1}{\|\mathbf{w}\|_1} w_i \sum_{i=1}^n f(\xi_i)$ and $v_{\mathbf{w}}(f) = P_{\mathbf{w}}(f) - P(f)$ where $0 \leq w_i \leq 1$ for all $i = 1, \ldots, n$ and $\|\mathbf{w}\|_1 = \sum_{i=1}^n w_i > 0$. Given a collection $\mathcal{F}$ of P-integrable functions f, our purpose is to control the expectation of $\sup_{f\in\mathcal{F}} v_{\mathbf{w}}(f)$ or $\sup_{f\in\mathcal{F}} -v_{\mathbf{w}}(f)$.

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 σ 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:

$$\sqrt{\|\mathbf{w}\|_1}\max(\mathbb{E}[\sup_{f\in\mathcal{F}} v_{\mathbf{w}}(f_0 - f)], \mathbb{E}[\sup_{f\in\mathcal{F}} v_{\mathbf{w}}(f - f_0)]) \leq 12\varphi(\sigma).$$

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

C Inequalities for $U_{\mathbf{w}}(r)$ and $\widetilde{U}_{\mathbf{w}}(r)$

In this section, we first show the classical symmetrization and randomization tricks for the degenerated weighted U-statistics $U_{\mathbf{w}}(r)$ and the degenerated part $\widetilde{U}_{\mathbf{w}}(r)$. Then we establish general exponential inequalities for weighted Rademacher chaos. This result is generalized from (Clémençon, 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 $\mathbb{L}_\infty$ 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:*

$$\begin{aligned}\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j} f_{i,j}(X_i, X_j)|^q] \\ \leq 4^q \mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j} f_{i,j}(X_i, X_j')|^q]\end{aligned} \tag{22}$$

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 $(w_{i,j})_{(i,j)\in E}$ is symmetric, then the inequality can be reversed, that is,

$$\begin{aligned}\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in\mathbb{E}} w_{i,j} f_{i,j}(X_i, X_j')|^q] \\ \leq 4^q \mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j} f_{i,j}(X_i, X_j)|^q]\end{aligned} \tag{23}$$

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 $q \geq 1$,*

$$\begin{aligned}\mathbb{E}[\sup_{f\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j} f(X_i, X_j')|^q] \\ \leq 4^q \mathbb{E}[\sup_{f\in\mathcal{F}} |\sum_{(i,j)\in E} \sigma_i\sigma_j' w_{i,j} f(X_i, X_j')|^q]\end{aligned}$$

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 $\{0,1\}$, $(\tilde{Y}_{i,j})_{(i,j)\in E}$, conditionally independent given the X_i''s and the X_i's and such that $P[\tilde{Y}_{i,j} = 1 \mid X_i, X_j'] = \eta(X_i, X_j')$. We have for all $q \geq 1$,*

$$\begin{aligned}\mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in E} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j})|^q] \\ \leq 4^q \mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in E} w_{i,j} \tilde{h}_r(X_i, X_j', \tilde{Y}_{i,j})|^q]\end{aligned} \tag{24}$$

and the inequality can be reversed,

$$\begin{aligned}\mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in\mathbb{E}} w_{i,j} \tilde{h}_r(X_i, X_j', \tilde{Y}_{i,j})|^q] \\ \leq 4^q \mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in E} w_{i,j} \tilde{h}_r(X_i, X_j, Y_{i,j})|^q]\end{aligned} \tag{25}$$

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 $q \geq 1$,*

$$\begin{aligned}\mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in E} w_{i,j} \tilde{h}_r(X_i, X_j', \tilde{Y}_{i,j})|^q] \\ \leq 4^q \mathbb{E}[\sup_{r\in R} |\sum_{(i,j)\in E} \sigma_i\sigma_j' w_{i,j} \tilde{h}_r(X_i, X_j', \tilde{Y}_{i,j})|^q]\end{aligned}$$

Lemma 5 and 6 are applications of Theorem 3.1.1 and Theorem 3.5.3 of (de la Peña and Giné 1999) 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 $\widetilde{U}_{\mathbf{w}}(r)$.

Theorem 6 (moment inequality)**.** *Let $X, X_1, \ldots, X_n$ be i.i.d. random variables and let $\mathcal{F}$ be a class of kernels. Consider*

a weighted Rademacher chaos Z_σ of order 2 on the graph $G = (V, E)$ indexed by $\mathcal{F}$,

$$Z = \sup_{f\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j} f(X_i, X_j)|$$

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

$$Z_\sigma = \sup_{f\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j}\sigma_i\sigma_j f(X_i, X_j)|$$

$$U_\sigma = \sup_{f\in\mathcal{F}} \sup_{\alpha:\|\alpha\|_2\le 1} \sum_{(i,j)\in\mathbb{E}} w_{i,j}\sigma_i\alpha_j f(X_i, X_j)$$

$$M = \sup_{f\in\mathcal{F}, k=1,\dots,n} |\sum_{i:(i,k)\in E} w_{i,k}\sigma_i f(X_i, X_k)|$$

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

$$\begin{aligned} &P[Z \ge C\mathbb{E}[Z_\sigma] + t] \\ &\quad\le \exp\left(-\frac{1}{C}\min\Big((\frac{t}{\mathbb{E}[U_\sigma]})^2, \frac{t}{\mathbb{E}[M] + F\|\mathbf{w}\|_2}, \right. \\ &\qquad \left. (\frac{t}{\|\mathbf{w}\|_{\max} F})^{2/3}, \sqrt{\frac{t}{\|\mathbf{w}\|_\infty F}}\Big)\right) \end{aligned}$$

where $\|\mathbf{w}\|_{\max} = \max_i \sqrt{\sum_{j:(i,j)\in E} w_{i,j}^2}$.

If the hypothesis set $\mathcal{F}$ is a subset of $\mathbb{L}_\infty(\mathcal{X}^2)$ (upper bounds on the uniform covering number with $\mathbb{L}_\infty$ 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 $N_\infty(\mathcal{F}, \epsilon)$ since all these Rademacher random variables satisfy the Khinchine inequality (see Section F). Following the metric entropy inequality for Khinchine-type processes (see Section F), it is easy to get the following Corollary.

Corollary 1. *With the same setting of Theorem 6, if $\mathcal{F} \subset \mathbb{L}_\infty(\mathcal{X}^2)$, we have for any $\delta < 1/e$,*

$$P[Z \le \kappa] \ge 1 - \delta$$

where

$$\begin{aligned} \kappa \le & C\left(\|\mathbf{w}\|_2 \int_0^{2F} \log N_\infty(\mathcal{F}, \epsilon) d\epsilon \right. \\ &+ \max\Big(\|\mathbf{w}\|_2 \log(1/\delta) \int_0^{2F} \sqrt{\log N_\infty(\mathcal{F}, \epsilon)} d\epsilon, \\ &\left. (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max}, (\log(1/\delta))^2\|\mathbf{w}\|_\infty \Big)\right). \end{aligned}$$

with a universal constant C.

D Proofs of the main track

D.1 Proof of Lemma 1

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

$$P[\sup_{r\in R} |U_{\mathbf{w}}(r)| \le \kappa] \ge 1 - \delta$$

where

$$\begin{aligned} \kappa \le & \frac{C_1}{\|\mathbf{w}\|_1}\left(\|\mathbf{w}\|_2 \int_0^1 \log N_\infty(R, \epsilon) d\epsilon \right. \\ &+ \max\Big(\|\mathbf{w}\|_2 \log(1/\delta) \int_0^1 \sqrt{\log N_\infty(R, \epsilon)} d\epsilon, \\ &\left. (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max}, (\log(1/\delta))^2\|\mathbf{w}\|_\infty \Big)\right). \end{aligned}$$

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

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

which can be handled by the by-products of Theorem 6. More specifically, using (42) and (43) combined with the arguments in Corollary 1 and Assumption 1 will gives the second inequality for $\sup_{r\in R}|\widetilde{U}_{\mathbf{w}}(r)|$.

D.2 Proof of Lemma 2

For any function $r \in \mathcal{R}$, observe first that

$$\begin{aligned} &\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1] \\ &\quad = \mathbb{E}[\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1, X_2] \mid X_1] \\ &\quad = \mathbb{E}[|1 - 2\eta(X_1, X_2)| \mathbb{1}_{r(X_1,X_2)\ne r^*(X_1,X_2)} \mid X_1] \end{aligned}$$

Then observing that

$$|1 - 2\eta(X_1, X_2)|^2 \le |1 - 2\eta(X_1, X_2)|$$

almost sure, and combining with Jensen inequality, we have

$$\begin{aligned} &\mathrm{Var}[\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1]] \\ &\quad \le \mathbb{E}[(\mathbb{E}[q_r(X_1, X_2, Y_{1,2}) \mid X_1])^2] \\ &\quad \le \mathbb{E}[|1 - 2\eta(X_1, X_2)| \mathbb{1}_{r(X_1,X_2)\ne r^*(X_1,X_2)}] \\ &\quad = \Lambda(r). \end{aligned}$$

D.3 Proof of Lemma 3

First, we introduce some notations of weighted ERM of the i.i.d. case. We denote $\{\bar{w}_i = \sum_{j:(i,j)\in E} w_{i,j} : i = 1, \dots, n\}$ the weights on vertices and introduce the "loss function"

$$\gamma(r, X) = 2h_r(X) + \Lambda(r)$$

and the weighted empirical loss of vertices

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

Define centered empirical process

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

and the pseudo-distance

$$d(r, r') = \frac{\sqrt{\|\bar{\mathbf{w}}\|_1}}{\|\bar{\mathbf{w}}\|_2}\left(\mathbb{E}[(\gamma(r, X) - \gamma(r', X))^2]\right)^{1/2}$$

for every $r, r' \in R$. Let ϕ be

$$\phi(\sigma) = 12\int_0^{\sigma}(\log N_\infty(R, \epsilon^2))^{1/2}d\epsilon. \tag{26}$$

From the definition of "loss function" γ, we have the excess risk of r is

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

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

$$d(r, r^*) \leq \frac{\sqrt{\|\bar{\mathbf{w}}\|_1}}{\|\bar{\mathbf{w}}\|_2}\sqrt{5\bar{\ell}(r, r^*)}$$

which implies that the modulus of continuity ψ can be taken as

$$\psi(\epsilon) = \sqrt{5}\epsilon. \tag{27}$$

Then from Lemma 4, we have

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

provided that $\phi(\sigma)/3 \leq \sqrt{\|\bar{\mathbf{w}}\|_1}\sigma^2$. It remains to bound the excess risk of $r_{\mathbf{w}}$ by the tight bounds for weighted ERM on i.i.d. examples by Theorem 5. For any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$\begin{aligned} L(r_{\mathbf{w}}) - L^* \leq 2(&\inf_{r\in R} L(r) - L^*) + 2\rho \\ &+ K\log(1/\delta)\epsilon_*^2) \end{aligned} \tag{28}$$

where C is a universal constant and ϵ_* is the unique positive solution of the equation

$$\sqrt{\|\bar{\mathbf{w}}\|_1}\epsilon_*^2 = \phi(\psi(\epsilon_*)).$$

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

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

which completes the proof.

D.4 Proof of Theorem 1

Let us consider the Hoeffding decomposition (5) of $\Lambda_{\mathbf{w}}(r)$ that is minimized over $r \in R$. The idea of this proof is that the degenerate part $U_{\mathbf{w}}(r)$ and $\widetilde{U}_{\mathbf{w}}(r)$ can be bounded by the moment inequality for weighted U-processes. Therefore, $r_{\mathbf{w}}$ is an approximate minimizer of $T_{\mathbf{w}}(r)$, which can be handled by Lemma 3.

Let A be the event on which

$$\sup_{r\in R}|U_{\mathbf{w}}(r)| \leq \kappa_A,$$

where

$$\kappa_A = \frac{C_A}{1-\beta}\max\Big(\frac{\|\mathbf{w}\|_2\log(1/\delta)}{\|\mathbf{w}\|_1}, \frac{\|\mathbf{w}\|_{\max}(\log(1/\delta))^{3/2}}{\|\mathbf{w}\|_1}, \frac{\|\mathbf{w}\|_\infty(\log(1/\delta))^2}{\|\mathbf{w}\|_1}\Big)$$

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

$$\sup_{r\in R}|\widetilde{U}_{\mathbf{w}}(r)| \leq \kappa_B.$$

where

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

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

By (5), it is clear that, on $A \cap B$, $r_{\mathbf{w}}$ is a ρ-minimizer of $T_{\mathbf{w}}(r)$ over $r \in R$ in the sense that the value of this latter quantity at its minimum is at most $(\kappa_A + \kappa_B)$ smaller than at $r_{\mathbf{w}}$. Then, from Lemma 3, with probability at least $1 - \delta/2$, $r_{\mathbf{w}}$ is a $(\kappa_A + \kappa_B)$-minimizer of $T_{\mathbf{w}}(r)$, which the result follows.

E Proofs in the appendix

E.1 Proof of Theorem 3

We write, for all $r \in R$,

$$\begin{aligned} L(r) - L_{\mathbf{w}}(r) &\leq \sup_{r\in R}\left[\mathbb{E}[\ell(r, z)] - \frac{1}{\|\mathbf{w}\|_1}\sum_{i=1}^{m}w_i\ell(r, z_i)\right] \\ &\leq \sum_{j=1}^{J}p_j\left[\sup_{r\in R}\sum_{i\in\mathcal{C}_j}\frac{w_{k_j^i}}{\|\mathbf{w}\|_1}(\mathbb{E}[\ell(r, z)] - \ell(r, z_{k_j^i}))\right]. \end{aligned} \tag{29}$$

Now, consider, for each j,

$$f_j(S_{\mathcal{C}_j}) = \sup_{r\in R}\sum_{i\in\mathcal{C}_j}\frac{w_{k_j^i}}{\|\mathbf{w}\|_1}(\mathbb{E}[\ell(r, z)] - \ell(r, z_{k_j^i})).$$

Let f is defined by, for all training set S, $f(S) = \sum_{j=1}^{J}p_jf_j(S_{\mathcal{C}_j})$, then f satisfies the conditions of Theorem 2 with, for any $i \in \{1, \ldots, m\}$, $\beta_i \leq w_i/\|\mathbf{w}\|_1$. Therefore, we can claim that, with probability at least $1 - \delta$,

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

$$+\sqrt{\chi^*(D_G)\log(1/\delta)}\frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_1}$$

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 $\mathfrak{R}^*_{\mathbf{w}}(R,S)$ which completes the proof.

E.2 Proof of Theorem 4

We use a auxiliary random variable

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

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

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

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

$$\widetilde{Z}-Z_k \leq \frac{1}{2b}w_i(f_0(X_k)-\mathbb{E}[f_0(X_k)]) \leq 1 \; a.s.,$$

$$\widetilde{Z}-Z_k-Z'_k \geq 0$$

and

$$\mathbb{E}[Z'_k]=0.$$

The first inequality is derived from $w_i \leq 1$ and $\sup_{f\in\mathcal{F},X\in\mathcal{X}} f(X)-\mathbb{E}[f(X)] \leq 2b$. Also, we have

$$\begin{aligned}(n-1)\widetilde{Z} &= \sum_{k=1}^{n}\frac{1}{2b}\sum_{i\neq k} w_i(f_0(X_i)-\mathbb{E}[f_0(X_i)]) \\ &\leq \sum_{k=1}^{n} Z_k,\end{aligned}$$

and

$$\begin{aligned}\sum_{k=1}^{n}\mathbb{E}^k_n[Z'^2_k] &= \frac{1}{2b}\sum_{k=1}^{n}\mathbb{E}[w_i^2(f_k(X_k)-\mathbb{E}[f_k(X_k)])^2] \\ &\leq \frac{1}{4b^2}\|\mathbf{w}\|_2^2\sup_{f\in\mathcal{F}}\mathrm{Var}[f(X)] \\ &\leq \frac{1}{4b^2}\|\mathbf{w}\|_2^2\sigma^2.\end{aligned}$$

where $\sigma^2 \geq \sup_{f\in\mathcal{F}}\mathrm{Var}[f(X)]$. Notice that we use the fact the X_i have identical distribution. Applying Theorem 1 of (Bousquet 2002) with $v = 2\mathbb{E}[\widetilde{Z}]+\frac{\|\mathbf{w}\|_2^2}{4b^2}\sigma^2$ will give

$$P[\widetilde{Z}-\mathbb{E}[\widetilde{Z}] \geq \sqrt{2vx}+\frac{x}{3}] \leq e^{-x},$$

and then

$$\begin{aligned}P[\frac{\|\mathbf{w}\|_1}{2b}(Z-\mathbb{E}[Z]) \geq \sqrt{2x(\frac{\|\mathbf{w}\|_1}{b}\mathbb{E}[Z]+\frac{\|\mathbf{w}\|_2^2}{4b^2}\sigma^2)} \\ +\frac{x}{3}] \leq e^{-x}\end{aligned}$$

which proves the inequality.

E.3 Proof of Theorem 5

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 $P[\gamma(\cdot,r_k)]$ that tends to $P[\gamma(\cdot,r)]$ as k tends to infinity. Denote the bias term of loss $\bar{\ell}(r^*,R) = \inf_{r\in R}\bar{\ell}(r^*,r)$. Hence, $\bar{\ell}(r^*,R) = \bar{\ell}(r^*,R')$, which implies that there exists some point $\pi(r^*)$ (which may depend on ϵ_*) such that $\pi(r^*)\in R'$ and

$$\bar{\ell}(r^*,\pi(r^*)) \leq \bar{\ell}(r^*,R)+\epsilon_*^2. \tag{30}$$

We start from the identity

$$\begin{aligned}\bar{\ell}(r^*,\hat{r}) = \bar{\ell}(r^*,\pi(r^*))+\gamma_{\mathbf{w}}(\hat{r})-\gamma_{\mathbf{w}}(\pi(r^*)) \\ +\overline{\gamma_{\mathbf{w}}}(\pi(r^*))-\overline{\gamma_{\mathbf{w}}}(\hat{r})\end{aligned}$$

which, by definition of $\hat{r}$, implies that

$$\bar{\ell}(r^*,\hat{r}) \leq \rho+\bar{\ell}(r,\pi(r^*))+\overline{\gamma_{\mathbf{w}}}(\pi(r^*))-\overline{\gamma_{\mathbf{w}}}(\hat{r}).$$

Let $x=\sqrt{K'y}\epsilon_*$, where K' is a constant to be chosen later such that $K'\geq 1$ and

$$V_x = \sup_{r\in R}\frac{\overline{\gamma_{\mathbf{w}}}(\pi(r^*))-\overline{\gamma_{\mathbf{w}}}(r)}{\bar{\ell}(r^*,r)+\epsilon^*+x^2}.$$

Then,

$$\bar{\ell}(r^*,\hat{r}) \leq \rho+\bar{\ell}(r^*,\pi(r^*))+V_x(\bar{\ell}(r^*,\hat{r})+x^2+\epsilon_*^2)$$

and therefore, on the event $V_x<1/2$, one has

$$\bar{\ell}(r^*,\hat{r}) \leq 2(\rho+\bar{\ell}(r^*,\pi(r^*)))+\epsilon_*^2+x^2,$$

yielding

$$\begin{aligned}P[\bar{\ell}(r^*,\hat{r}) \leq 2(\rho+\bar{\ell}(r^*,\pi(r^*)))+3\epsilon_*^2+x^2] \\ \leq P[V_x\geq\frac{1}{2}].\end{aligned} \tag{31}$$

Since $\bar{\ell}$ is bounded by 1, we may always assume x (and thus ϵ_*) to be not larger than 1. Assuming that $x\leq 1$, it remains to control the variable V_x via Theorem 4. In order to use Theorem 4, we first remark that, by Condition 3,

$$V_x = \sup_{r\in R'}\frac{\overline{\gamma_{\mathbf{w}}}(\pi(r^*))-\overline{\gamma_{\mathbf{w}}}(r)}{\bar{\ell}(r^*,r)+\epsilon^*+x^2}$$

which means that we indeed have to deal with a countably indexed empirical process. Note that the triangle inequality implies via (17), (30) and (18) that

$$\begin{aligned}(\mathrm{Var}[\gamma(r,\cdot)-\gamma(\pi(r^*),\cdot)])^{\frac{1}{2}} &\leq d(r^*,r)+d(r^*,\pi(r^*)) \\ &\leq 2\frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi(\sqrt{\bar{\ell}(r^*,r)+\epsilon_*^2})\end{aligned} \tag{32}$$

Since γ takes its values in $[0,1]$, introducing the functions $\psi_1 = \min(1, 2\psi)$ and, we derive from (32) that

$$\sup_{r\in R} \mathrm{Var}[\frac{\overline{\gamma_{\mathbf{w}}}(\pi(r^*)) - \overline{\gamma_{\mathbf{w}}}(t)}{\bar{\ell}(r^*,t) + \epsilon^* + x^2}] \leq \sup_{\epsilon\geq 0} \frac{(\frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi_1(\epsilon))^2}{(\epsilon^2+x^2)^2}$$
$$\leq \frac{\|\mathbf{w}\|_1}{\|\mathbf{w}\|_2^2 x^2} \sup_{\epsilon\geq 0}(\frac{\psi_1(\epsilon)}{\max(\epsilon,x)})^2.$$

Now the monotonicity assumptions on ψ imply that either $\psi(\epsilon) \leq \psi(x)$ if $x \geq \epsilon$ or $\psi(\epsilon)/\epsilon \leq \psi(x)/x$ if $x \leq \epsilon$. Hence, one has in any case $\psi(\epsilon)/(\max(\epsilon,x)) \leq \psi(x)/x$, which finally yields

$$\sup_{r\in R} \mathrm{Var}[\frac{\overline{\gamma_{\mathbf{w}}}(\pi(r^*)) - \overline{\gamma_{\mathbf{w}}}(r)}{\bar{\ell}(r^*,r) + \epsilon^* + x^2}] \leq \frac{\|\mathbf{w}\|_1 \psi_1^2(x)}{\|\mathbf{w}\|_2^2 x^4}.$$

On the other hand, since γ takes its values in $[0,1]$, we have

$$\sup_{r\in R} \| \frac{\gamma(r,\cdot) - \gamma(\pi(r^*),\cdot)}{\bar{\ell}(r^*,r) + x^2} \|_\infty \leq \frac{1}{x^2}.$$

We can therefore apply Theorem 4 with $v = \psi_1^2(x)x^{-4}$ and $b = x^{-2}$, which gives that, on a set Ω_y with probability larger than $1 - \exp(-y)$, the inequality

$$V_x < \mathbb{E}[V_x] + \sqrt{\frac{2y(\psi_1^2(x)x^{-2} + 4\mathbb{E}[V_x])}{\|\mathbf{w}\|_1 x^2}} + \frac{2y}{3\|\mathbf{w}\|_1 x^2}. \tag{33}$$

Now since ϵ_* is assumed to be not larger than 1, one has $\psi(\epsilon_*) \geq \epsilon_*$ and therefore, for every $\sigma \geq \psi(\epsilon_*)$, the following inequality derives from the definition of ϵ_* by monotonicity:

$$\frac{\phi(\sigma)}{\sigma^2} \leq \frac{\phi(\psi(\epsilon_*))}{w^2(\epsilon_*)} \leq \frac{\phi(\psi(\epsilon_*))}{\epsilon_*^2} = \sqrt{\|\mathbf{w}\|_1}.$$

Thus, (19) holds for every $\sigma \geq \psi(\epsilon_*)$. In order to control $\mathbb{E}[V_x]$, we intend to use Lemma A.5 of (Massart and Nédélec 2006). For every $r \in R'$, we introduce $a^2(r) = \max(\bar{\ell}(r^*,\pi(r^*)), \bar{\ell}(r^*,r))$. Then by (30), $\bar{\ell}(r^*,r) \leq a^2(r) \leq \bar{\ell}(r^*,r) + \epsilon_*^2$. Hence, we have, on the one hand, that

$$\mathbb{E}[V_x] \leq \mathbb{E}[\sup_{r\in R'} \frac{\overline{\gamma_{\mathbf{w}}}(\pi(r^*)) - \overline{\gamma_{\mathbf{w}}}(r)}{a^2(r) + x^2}].$$

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

$$\mathbb{E}[\sup_{r\in R', a(r)\leq\epsilon} \overline{\gamma_{\mathbf{w}}}(\pi(r^*)) - \overline{\gamma_{\mathbf{w}}}(r)]$$
$$\leq \mathbb{E}[\sup_{r\in R', \bar{\ell}(r^*,r)\leq\epsilon^2} \overline{\gamma_{\mathbf{w}}}(\pi(r^*)) - \overline{\gamma_{\mathbf{w}}}(r)].$$

Now by (30) if there exists some $r \in R'$ such that $\bar{\ell}(r^*,r) \leq \epsilon^2$, then $\bar{\ell}(r^*,\pi(r^*)) \leq \epsilon^2 + \epsilon_*^2 \leq 2\epsilon^2$ and therefore, by assumption (18) and monotonicity of $\theta \to \psi(\theta)/\theta$, $d(\pi(r^*),r) \leq 2\frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi(\sqrt{2}\epsilon) \leq 2\sqrt{2}\frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi(\epsilon)$, then $\frac{\|\mathbf{w}\|_2}{\sqrt{\|\mathbf{w}\|_1}} d(\pi(r^*),r) \leq 2\sqrt{2}\psi(\epsilon)$. Thus, we derive from (19) that, for every $\epsilon \geq \epsilon_*$,

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

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

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

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

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

Thus, using the monotonicity of $\theta \to \phi(\psi(\theta))/\theta$, and the definition of ϵ_*, we derive that

$$\mathbb{E}[V_x] \leq \frac{8\sqrt{2}\phi(\psi(\epsilon_*))}{\sqrt{\|\mathbf{w}\|_1}x\epsilon_*} = \frac{8\sqrt{2}\epsilon_*}{x} \leq \frac{8\sqrt{2}}{\sqrt{K'y}} \leq \frac{8\sqrt{2}}{\sqrt{K'}}, \tag{34}$$

provided that $x \geq \epsilon_*$, which holds since $K' \geq 1$. Now, the monotonicity of $\theta \to \psi_1(\theta)/\theta$ implies that $x^{-2}\psi_1^2(x) \leq \epsilon_*^{-2}\psi_1^2(\epsilon_*)$, but since $\phi(\theta)/\theta \geq \phi(1) \geq 1$ for every $\theta \in [0,1]$, we derive from (20) and the monotonicity of ϕ and $\theta \to \phi(\theta)/\theta$ that

$$\frac{\psi_1^2(\epsilon_*)}{\epsilon_*^2} \leq \frac{\phi^2(\psi_1(\epsilon_*))}{\epsilon_*^2} \leq \frac{\phi^2(2\psi(\epsilon_*))}{\epsilon_*^2} \leq 4\frac{\phi^2(\psi(\epsilon_*))}{\epsilon_*^2}$$

and, therefore, $x^{-2}\psi_1^2(x) \leq 4\|\mathbf{w}\|_1\epsilon_*^2$. Plugging this inequality together with (34) into (33) implies that, on the set Ω_y,

$$V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \sqrt{\frac{2y(4\|\mathbf{w}\|_1\epsilon_*^2 + 32/\sqrt{K'})}{\|\mathbf{w}\|_1 x^2}} + \frac{2y}{3\|\mathbf{w}\|_1 x^2}.$$

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

$$V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \sqrt{\frac{8(1 + 4(\|\mathbf{w}\|_1\epsilon_*^2\sqrt{K'})^{-1})}{K'}} + \frac{2}{3\|\mathbf{w}\|_1 K'\epsilon_*^2}.$$

Taking into account that $\phi(\psi(\theta)) \geq \phi(\min(1,\psi(\theta))) \geq \theta$ for every $\theta \in [0,1]$, we deduce from the definition of ϵ_* that $\|\mathbf{w}\|_1\epsilon_*^2 \geq 1$ and, therefore, the preceding inequality becomes, on Ω_y,

$$V_x < \frac{8\sqrt{2}}{\sqrt{K'}} + \sqrt{\frac{8(1 + 4/\sqrt{K'})}{K'}} + \frac{2}{3K'}.$$

Hence, choosing K' as a large enough numerical constant guarantee that $V_x < 1/2$ on Ω_y and, therefore, (31) yields

$$P[\bar{\ell}(r^*,\hat{r}) \leq 2(\rho + \bar{\ell}(r^*,\pi(r^*))) + 3\epsilon_*^2 + x^2]$$
$$\leq P[\Omega_y^c]$$
$$\leq e^{-y}.$$

We get the required probability bound (5) by setting $K = K' + 3$.

E.4 Proof of Lemma 4

We first perform the control of $\mathbb{E}[\sup_{f\in\mathcal{F}} v_{\mathbf{w}}(f - f_0)]$. For simplicity, we denote $H_\infty(\mathcal{F},\epsilon) = \log N_\infty(\mathcal{F},\epsilon)$. For any integer j, we set $\sigma_j = \sigma 2^{-j}$ and $H_j = H_\infty(\mathcal{F},\sigma_j^2)$. By definition of $H_j = H_\infty(\mathcal{F},\sigma_j^2)$, for any integer $j \geq 1$, we

can define a mapping Π_j from $\mathcal{F}$ to some finite collection of functions such that

$$\log \#\{\Pi_j \mathcal{F}\} \leq H_j \tag{35}$$

and

$$\Pi_j f \leq f \text{ with } P(f - \Pi_j f) \leq \sigma_j^2, \forall f \in \mathcal{F}. \tag{36}$$

For $j = 0$, we choose Π_0 to be identically equal to f_0. For this choice of Π_0, we still have

$$P(|f - \Pi_0 f|) = P[|f - f_0|] \leq \sigma_0^2 = \sigma \tag{37}$$

for every $f \in \mathcal{F}$. Furthermore, since we may always assume that the extrmities of the balls used to cover $\mathcal{F}$ take their values in $[0, 1]$, we also have for every integer j that

$$0 \leq \Pi_j f \leq 1.$$

Noticing that since $u \to H_\infty(\mathcal{F}, u^2)$ is nonincreasing,

$$H_1 \leq \sigma_1^{-2} \varphi^2(\sigma),$$

and under the condition $4\varphi(\sigma) \leq \sigma^2 \sqrt{\|\mathbf{w}\|_1}$, one has $H_1 \leq \sigma_1^2 \|\mathbf{w}\|_1$. Thus, since $j \to H_j \sigma_j^{-2}$ increases to infinity, the set $\{j \geq 0 : H_j \leq \sigma_j^2 \|\mathbf{w}\|_1\}$ is a nonvoid interval of the form

$$\{j \geq 0 : H_j \leq \sigma_j^2 \|\mathbf{w}\|_1\} = [0, J],$$

with $J \geq 1$. For every $f \in \mathcal{F}$, starting from the decomposition

$$\begin{aligned} -v_{\mathbf{w}}(f) = \sum_{j=0}^{J-1} v_{\mathbf{w}}(\Pi_j f) - v_{\mathbf{w}}(\Pi_{j+1} f) \\ + v_{\mathbf{w}}(\Pi_J f) - v_{\mathbf{w}}(f), \end{aligned}$$

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

$$-v_{\mathbf{w}}(f) = \sum_{j=0}^{J-1} v_{\mathbf{w}}(\Pi_j f) - v_{\mathbf{w}}(\Pi_{j+1} f) + \sigma_J^2$$

and, therefore,

$$\begin{aligned} &\mathbb{E}[\sup_{f \in \mathcal{F}} [-v_{\mathbf{w}}(f)]] \\ &\leq \sum_{j=0}^{J-1} \mathbb{E}[\sup_{f \in \mathcal{F}} v_{\mathbf{w}}(\Pi_j f) - v_{\mathbf{w}}(\Pi_{j+1} f)] + \sigma_J^2. \end{aligned} \tag{38}$$

Now, it follows from (36) and (37) 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, (35) 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, Ramon, and Guo 2014, Lemma 16) that

$$\begin{aligned} &\sqrt{\|\mathbf{w}\|_1} \mathbb{E}[\sup_{f \in \mathcal{F}} [v_{\mathbf{w}}(\Pi_j f) - v_{\mathbf{w}}(\Pi_{j+1} f)]] \\ &\leq 2[\sigma_{j+1} \sqrt{5H_{j+1}} + \frac{1}{3\sqrt{\|\mathbf{w}\|_1}} H_{j+1}] \end{aligned}$$

because $w_i \leq 1, \forall i \in 1, \ldots, n$, and (38) becomes

$$\begin{aligned} &\sqrt{\|\mathbf{w}\|_1} \mathbb{E}[\sup_{f \in \mathcal{F}} -v_{\mathbf{w}}(f)] \\ &\leq 2 \sum_{j=1}^{J} [\sigma_j \sqrt{5H_j} + \frac{1}{3\sqrt{\|\mathbf{w}\|_1}} H_j] + 4\sqrt{\|\mathbf{w}\|_1} \sigma_{J+1}^2. \end{aligned} \tag{39}$$

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,

$$4\sqrt{\|\mathbf{w}\|_1} \sigma_{J+1}^2 \leq 4\sigma_{J+1} \sqrt{H_{j+1}}.$$

Hence, plugging these inequalities in (39) yields

$$\sqrt{\|\mathbf{w}\|_1} \mathbb{E}[\sup_{f \in \mathcal{F}} -v_{\mathbf{w}}(f)] \leq 6 \sum_{j=1}^{J+1} \sigma_j \sqrt{H_j},$$

and the result follows. The control of $\mathbb{E}[\sup_{f \in \mathcal{F}} v_{\mathbf{w}}(f - f_0)]$ can be performed analogously.

E.5 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 1999).

For any random variable X, we denote by $\mathcal{L}(X)$ its distribution. We denote by Σ (respectively Σ') the sigma-filed generated by $\{X_1, \ldots, X_n\}$ (respectively $\{X_1', \ldots, X_n'\}$). Let $(Y_{i,j}')_{(i,j) \in E}$ be Bernoulli random variables such that $P[Y_{i,j}' = 1 \mid \Sigma, \Sigma'] = \eta(X_i', X_j')$. Let $(\sigma_i)_{i=1}^n$ be independent Rademacher variables and define:

$$\begin{aligned} Z_i &= X_i \text{ if } \sigma_i = 1, \text{ and } X_i' \text{ otherwise,} \\ Z_i' &= X_i' \text{ if } \sigma_i = 1, \text{ and } X_i' \text{ otherwise.} \end{aligned}$$

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:

$$\begin{aligned} &\mathcal{L}(X_1, \ldots, X_n, X_1', \ldots, X_n') = \\ &\mathcal{L}(Z_1, \ldots, Z_n, Z_1', \ldots, Z_n') \end{aligned} \tag{40}$$

and

$$\mathcal{L}(X_1, \ldots, X_n) = \mathcal{L}(Z_1, \ldots, Z_n). \tag{41}$$

Let $\{\tilde{Y}_{i,j}\}_{(i,j) \in E}$ be Bernoulli random variables such that $P[\tilde{Y}_{i,j} = 1 \mid \Sigma, \Sigma'] = \eta(X_i, X_j')$ and define for $(i, j) \in E$,

$$\hat{Y}_{i,j} = \begin{cases} Y_{i,j} & \text{if } \sigma_i = 1 \text{ and } \sigma_j = -1 \\ Y_{i,j}' & \text{if } \sigma_i = -1 \text{ and } \sigma_j = 1 \\ \tilde{Y}_{i,j} & \text{if } \sigma_i = 1 \text{ and } \sigma_j = 1 \\ \tilde{Y}_{i,j} & \text{if } \sigma_i = -1 \text{ and } \sigma_j = -1 \end{cases}$$

Notice that for all f,

$$\begin{aligned}\mathbb{E}_\sigma[\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j})] \\ = \frac{1}{4}(\tilde{h}_r(X_i, X_j, Y_{i,j}) + \tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) \\ + \tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) + \tilde{h}_r(X'_i, X'_j, Y'_{i,j}))\end{aligned}$$

where $\mathbb{E}_\sigma$ denotes the expectation taken with respect to $\{\sigma_i\}_{i=1}^n$. Moreover, using

$$\mathbb{E}[\tilde{h}_r(X'_i, X'_j, Y'_{i,j}) \mid \Sigma] = 0$$

and (degenerated)

$$\mathbb{E}[\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) \mid \Sigma] = 0$$
$$\mathbb{E}[\tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) \mid \Sigma] = 0$$

we easily get

$$\tilde{h}_r(X_i, X_j, Y_{i,j}) = 4\mathbb{E}[\tilde{h}_r(Z_i, Z'_j, \hat{Y}_{i,j}) \mid \Sigma]$$

For all $q \geq 1$, we therefore have

$$\begin{aligned}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q] \\ = \mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} 4w_{i,j}\mathbb{E}[\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j}) \mid \Sigma]|^q] \\ \leq 4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j})|^q]\end{aligned}$$

derived from the facts that the supreme and $|x|^p (p \geq 1)$ are convex functions and the Jansen inequality. According to (40) and the fact that the distribution of $\hat{Y}_{i,j}$ only depends on the realization Z_i, Z'_j, i.e. $P[\hat{Y}_{i,j} \mid Z_i, Z'_j] = \eta(Z_i, Z'_j)$, we obtain

$$\begin{aligned}4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\hat{h}_r(Z_i, Z'_j, \tilde{Y}_{i,j})|^q] \\ = 4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j})|^q]\end{aligned}$$

which concludes the proof of (24).

By the symmetry of $\tilde{h}_r$ in the sense that $\tilde{h}_r(X_i, X_j, Y_{i,j}) = \tilde{h}_r(X_j, X_i, Y_{j,i})$, we have

$$\begin{aligned}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j)|^q] \\ = \mathbb{E}[\sup_{r\in R}|\frac{1}{2}\sum_{(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) \\ + \tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}))|^q] \\ = \mathbb{E}[\sup_{r\in R}|\frac{1}{2}\sum_{(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) \\ + \tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) + \tilde{h}_r(X_i, X_j, Y_{i,j}) \\ + \tilde{h}_r(X'_i, X'_j, Y'_{i,j}) \\ - \frac{1}{2}\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j}) \\ - \frac{1}{2}\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X'_i, X'_j, Y'_{i,j})|^q]\end{aligned}$$

(Triangle's Inequality and the convexity of $\sup_{r\in R}|\cdot|^q$)

$$\begin{aligned}\leq \frac{1}{2}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) \\ + \tilde{h}_r(X'_i, X_j, \tilde{Y}_{i,j}) + \tilde{h}_r(X_i, X_j, Y_{i,j}) \\ + \tilde{h}_r(X'_i, X'_j, Y'_{i,j})|^q] \\ + \frac{1}{4}\mathbb{E}[\sup_{r\in R}|2\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q] \\ + \frac{1}{4}\mathbb{E}[\sup_{r\in R}|2\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X'_i, X'_j, Y'_{i,j})|^q] \\ = \frac{1}{2}\mathbb{E}[\sup_{r\in R}|4\sum_{(i,j)\in E} w_{i,j}\mathbb{E}_\sigma[\tilde{h}_r(Z_i, Z_j, \tilde{Y}_{i,j})]|^q] \\ + \frac{1}{2}\mathbb{E}[\sup_{r\in R}|2\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q]\end{aligned}$$

(Jansen's Inequality and the convexity of $\sup_{r\in R}|\cdot|^q$)

$$\begin{aligned}\leq \frac{1}{2}\mathbb{E}[\sup_{r\in R}|4\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(Z_i, Z_j, \tilde{Y}_{i,j})|^q] \\ + \frac{1}{2}\mathbb{E}[\sup_{r\in R}|2\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q]\end{aligned}$$

(According to (41))

$$\begin{aligned}= \frac{1}{2}\mathbb{E}[\sup_{r\in R}|4\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q] \\ + \frac{1}{2}\mathbb{E}[\sup_{r\in R}|2\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q] \\ = 4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X_j, Y_{i,j})|^q]\end{aligned}$$

which (25) follows.

E.6 Proof of Lemma 8

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

$$\begin{aligned}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}''_{i,j})|^q] \\ (\tilde{h}_r \text{ is degenerated}) \\ = \mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}''_{i,j}) \\ - \mathbb{E}_{\Sigma''}[\tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j})])|^q]\end{aligned}$$

(Jansen's Inequality)

$$\leq \mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}''_{i,j}) - \tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j}))|^q]$$

$$= \mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sum_{i:(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}''_{i,j}) - \tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j}))|^q]$$

Let $(\sigma_i)_{i=1}^n$ be independent Rademacher variables, independent of Σ, Σ' and Σ'', then we have:

$$\mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sum_{i:(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) - \tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j}))|^q|\ \Sigma]$$

$$= \mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sigma_j\sum_{i:(i,j)\in E} w_{i,j}(\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j}) - \tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j}))|^q|\ \Sigma]$$

(Triangle's Inequality and the convexity of $\sup_{r\in R}|\cdot|^q$)

$$\leq 2^q\mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sigma_j\sum_{i:(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j})|^q|\ \Sigma] + 2^q\mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sigma_j\sum_{i:(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X''_j, \tilde{Y}''_{i,j}))|^q|\ \Sigma]$$

$$\leq 2^q\mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sigma_j\sum_{i:(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j})|^q|\ \Sigma]$$

and get

$$\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j})|^q] \leq 2^q\mathbb{E}[\sup_{r\in R}|\sum_{j=1}^{n}\sigma_j\sum_{i:(i,j)\in E} w_{i,j}\tilde{h}_r(X_i, X'_j, \tilde{Y}_{i,j})|^q]$$

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 togather.

E.7 Proof of Theorem 6

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

$$\mathbb{E}[\sup_f|\sum_{(i,j)\in E} w_{i,j}f(X_i, X_j)|^q] \leq 16^q\mathbb{E}[\sup_f|\sum_{(i,j)\in E} w_{i,j}\sigma_i\sigma'_j f(X_i, X'_j)|^q] \leq 64^q\mathbb{E}[\sup_f|\sum_{(i,j)\in E} w_{i,j}\sigma_i\sigma_j f(X_i, X_j)|^q]$$

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 $q \geq 2$,

$$(\mathbb{E}_\sigma[Z_\sigma^q])^{1/q} \leq \mathbb{E}_\sigma[Z_\sigma] + (E_\sigma[(Z_\sigma - \mathbb{E}_\sigma[Z_\sigma])_+^q])^{1/q} \leq \mathbb{E}_\sigma[Z_\sigma] + 3\sqrt{q}\mathbb{E}_\sigma U_\sigma + 4qB$$

where B is defined below

$$B = \sup_f \sup_{\alpha,\alpha':\|\alpha\|_2,\|\alpha'\|_2\leq 1} |\sum_{(i,j)\in E} w_{i,j}\alpha_i\alpha'_j f(X_i, X_j)|.$$

The second inequality above follows by Theorem 14 of (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

$$\mathbb{E}_\sigma[Z_\sigma^q] \leq 3^{q-1}(\mathbb{E}_\sigma[Z_\sigma]^q + 3^q q^{q/2}\mathbb{E}_\sigma[U_\sigma]^q + 4^q q^q B^q).$$

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

First term: $\mathbb{E}[\mathbb{E}_\sigma[Z_\sigma]^q]$. Using the symmetrization trick, we have

$$\mathbb{E}[\mathbb{E}_\sigma[Z_\sigma]^q] \leq 4^q\mathbb{E}[\mathbb{E}_\sigma[Z'_\sigma]^q]$$

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

$$A_k = \mathbb{E}_\sigma[\sup_f|\sum_{(i,j)\in E, i,j\neq k} w_{i,j}\sigma_i\sigma'_j f(X_i, X_j)|].$$

It is easy to see that $A_k \leq A$.

On the other hand, defining

$$R_k = \sup_f|\sum_{i:(i,k)\in E} w_{i,k}\sigma_i f(X_i, X_k)|,$$

$$M = \max_k R_k$$

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

$$\mathbb{E}[\mathbb{E}_\sigma[Z'_\sigma]^q] \leq 2^{q-1}(2^q(E[Z'_\sigma])^q + 5^q q^q\mathbb{E}[\mathbb{E}_\sigma[M]^q]) \quad (42)$$

To bound $\mathbb{E}[\mathbb{E}_\sigma[M]^q]$, observe that $\mathbb{E}_\sigma[M]$ is a conditional Rademacher average, for which Theorem 13

of (Boucheron et al. 2005) could be applied. Since $\max_k \sup_{f,i} w_{i,k} f(X_i, X_k) \leq \|\mathbf{w}\|_\infty F$, we have

$$\mathbb{E}[\mathbb{E}_\sigma[M]^q] \leq 2^{q-1}(2^q \mathbb{E}[M]^q + 5^q q^q \|\mathbf{w}\|_\infty^q F^q). \quad (43)$$

By undecoupling, we have $\mathbb{E}[Z'_\sigma] = \mathbb{E}[\mathbb{E}_{\sigma,\sigma'}[Z'_\sigma]] \leq \mathbb{E}[4\mathbb{E}_\sigma[Z_\sigma]] = 4\mathbb{E}[Z_\sigma]$. Collecting all terms, we have

$$\mathbb{E}[\mathbb{E}_\sigma[Z_\sigma]^q] \leq 64^q \mathbb{E}[Z_\sigma]^q + 160^q q^q \mathbb{E}[M]^q + 400^q \|\mathbf{w}\|_\infty^q F^q q^{2q}.$$

Second term: $\mathbb{E}[\mathbb{E}_\sigma[U_\sigma]^q]$.
By the Cauchy-Schwarz inequality we can observe that

$$\sup_{f,i} \sup_{\alpha:\|\alpha\|_2 \leq 1} \sum_{j:(i,j)\in E} w_{i,j} \alpha_j f(X_i, X_j) \leq \|\mathbf{w}\|_{\max} F.$$

Then similar to the bound of $\mathbb{E}[\mathbb{E}_\sigma[M]^q]$, we have

$$\mathbb{E}[\mathbb{E}_\sigma[U_\sigma]^q] \leq 2^{q-1}(2^q \mathbb{E}[U_\sigma]^q + 5^q q^q \|\mathbf{w}\|_{\max}^q F^q).$$

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

$$\mathbb{E}[B^q] \leq \|\mathbf{w}\|_2^q F^q.$$

Now it remains to simply put the pieces together to obtain

$$\begin{aligned}\mathbb{E}[\sup_f |\sum_{(i,j)\in E} w_{i,j} f(X_i, X_j)|^q] &\leq C(\mathbb{E}[Z_\sigma]^q + q^{q/2}\mathbb{E}[U_\sigma]^q + q^q \mathbb{E}[M]^q \\ &\quad + \|\mathbf{w}\|_\infty^q F^q q^{2q} + \|\mathbf{w}\|_{\max}^q F^q q^{3q/2} \\ &\quad + F^q \|\mathbf{w}\|_2^q q^q)\end{aligned}$$

for an appropriate constant C. In order to derive the exponential inequality, we use Markov's inequality $P[X \geq t] \leq t^{-q}\mathbb{E}[Z^q]$ and choose

$$q = C \min\Big((\frac{t}{\mathbb{E}[U_\sigma]})^2, \frac{t}{\mathbb{E}[M]}, \frac{t}{F\|\mathbf{w}\|_2}, (\frac{t}{\|\mathbf{w}\|_{\max} F})^{2/3}, \sqrt{\frac{t}{\|\mathbf{w}\|_\infty F}}\Big)$$

for an appropriate constant C.

E.8 Proof of Corollary 1

From Theorem 3, it is easy to know

$$\begin{aligned}\kappa =& C(\mathbb{E}[Z_\sigma] + \max(\mathbb{E}[U_\sigma]\sqrt{\log(1/\delta)}, \mathbb{E}[M]\log(1/\delta), \\ & \log(1/\delta)\|\mathbf{w}\|_2, (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max}, \\ & (\log(1/\delta))^2 \|\mathbf{w}\|_\infty)).\end{aligned} \quad (44)$$

An important character of these Rademacher processes in (44) is that they all satisfy the Khinchine inequality (46). For simplicity, we denote the weighted Rademacher processes of Z_σ by

$$\{z_\sigma(f) = \sum_{(i,j)\in E_<} w_{i,j}\sigma_i\sigma_j f(X_i, X_j), f \in \mathcal{F}\}.$$

where $E_< = \{(i,j) : \{i,j\} \in E, i < j\}$. Let $z'_\sigma = z_\sigma / \|\mathbf{w}\|_2$, following Theorem 7, we can easily have that z'_σ satisfies (46) with degree 2. Thus from Theorem 8, we have

$$\mathbb{E}_\sigma[\sup_{f,g} |z'_\sigma(f) - z'_\sigma(s)|] \leq K \int_0^D \log N(\mathcal{F}, d, \epsilon) d\epsilon. \quad (45)$$

where

$$\begin{aligned}d(f,g) &= (\mathbb{E}_\sigma[|z'_\sigma(f) - z'_\sigma(g)|^2])^{1/2} \\ &= \frac{1}{\|\mathbf{w}\|_2}(\mathbb{E}_\sigma[|\sum_{(i,j)\in E} w_{i,j}\sigma_i\sigma_j(f(X_i,X_j) \\ &\quad - g(X_i,X_j))|^2])^{1/2}.\end{aligned}$$

Recall that $\sup_f \|f\|_\infty = F$, we have $D = 2F$. Since this metirc function d is intractable, we need to convert it to the $\mathbb{L}_\infty$ metric. For all f, s, we have

$$\begin{aligned}d(f,g) &= \frac{1}{\|\mathbf{w}\|_2}(\mathbb{E}_\sigma[|\sum_{(i,j)\in E} w_{i,j}\sigma_i\sigma_j(f(X_i,X_j) \\ &\quad - g(X_i,X_j))|^2])^{1/2} \\ &\leq \mathbb{L}_\infty(f,g)\end{aligned}$$

The fact that if $\forall f, s \in \mathcal{F}, d_1(f,s) \leq d_2(f,s)$ then $N(\mathcal{F}, d_1, \epsilon) \leq N(\mathcal{F}, d_2, \epsilon)$ combined with (45) will give

$$\begin{aligned}\mathbb{E}[Z_\sigma] &= 2\mathbb{E}[\mathbb{E}_\sigma[\sup_f |z_\sigma(f)|]] \\ &\leq 2K\|\mathbf{w}\|_2 \int_0^{2F} \log N(\mathcal{F}, d, \epsilon) d\epsilon \\ &\leq 2K\|\mathbf{w}\|_2 \int_0^{2F} \log N_\infty(\mathcal{F}, \epsilon) d\epsilon\end{aligned}$$

where K is a universal constant. Similarly, we can also bound $\mathbb{E}[M]$ by $K\|\mathbf{w}\|_{\max} \int_0^{2F} \sqrt{\log N_\infty(\mathcal{F}, \epsilon)} d\epsilon$. For U_σ, let α^* be the (random) vector that maximizes U_σ and define

$$\{u_\sigma(f) = \sum_{i=1}^n \sigma_i \sum_{(i,j)\in\mathbb{E}} w_{i,j}\alpha_j^* f(X_i, X_j), f \in \mathcal{F}\}.$$

Clearly, u_σ satisfies the Kinchine inequality with degree 1. Also, we need to convert its metric distance,

$$\begin{aligned}&(\mathbb{E}_\sigma[|u_\sigma(f) - u_\sigma(g)|^2])^{1/2} \\ &= (\mathbb{E}_\sigma[|\sum_{(i,j)\in\mathbb{E}} w_{i,j}\sigma_i\alpha_j^*(f(X_i,X_j) \\ &\quad - g(X_i,X_j))|^2])^{1/2} \\ &\leq \|\mathbf{w}\|_2 \mathbb{L}_\infty(f,g)\end{aligned}$$

Thus, $\mathbb{E}[U_\sigma] \leq K\|\mathbf{w}\|_2 \int_0^{2F} \sqrt{\log N_\infty(\mathcal{F}, \epsilon)} d\epsilon$.

Pluging all these part into (44) will complete the corollary.

F Metric entropy inequality

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

Theorem 7 (Khinchine inequality for rademacher chaos, de la Peña and Giné (1999, 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

$$\gamma = (\frac{p-1}{q-1})^{1/2}.$$

Then, for all $d \geq 1$,

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

Theorem 8 (metric entropy inequality, Dembo, Karlin, and Zeitouni (1994, Proposition 2.6))**.** *If a process $\{Y_f : f \in \mathcal{F}\}$ satisfies*

$$(\mathbb{E}[|Y_f - Y_g|^p])^{1/p} \leq (\frac{p-1}{q-1})^{m/2}(\mathbb{E}[|Y_f - Y_g|^q])^{1/q}, \quad (46)$$

for $1 < q < p < \infty$ and some $m \geq 1$, and if

$$d(f,g) = (\mathbb{E}[|Y_f - Y_g|^2])^{1/2}, \quad (47)$$

there is a constant $K < \infty$ such that

$$\mathbb{E}[\sup_{f,g}|Y_f - Y_g|] \leq K\int_0^D (\log N(\mathcal{F}, d, \epsilon))^{m/2}d\epsilon. \quad (48)$$

where D is the d-diameter of $\mathcal{F}$.