Appendix

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

A Janson's Decomposition

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

Definition 1 (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 **??**. One can analyze this framework by the fractional coloring and the Janson's decomposition.

Definition 2 (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 fractional coloring of D_G and then analyze each set using the standard method for i.i.d. examples.

In particular, Theorem 1 (also see Usunier et al., 2006 Usunier et al., 2006, Theorem 2) provides a variation of McDiarmid's theorem using the Janson's decomposition. For simplicity, let $\mathbb{S} = \{z_i\}_{i=1}^m$ be the training examples drawn from $\mathcal{Z}^m$ and $\mathbb{S}_{\mathcal{C}_j}$ 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 $\mathbb{S}$.

Theorem 1. *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 $\mathbb{S}_{\mathcal{C}_j}$ and $\mathbb{S}^k_{\mathcal{C}_j}$ 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 et al., 2006), as shown in the following theorem.

Theorem 2. *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 et al., 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

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

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

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 (1) 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 assumption which can lead to "low-noise" condition (Tsybakov, 2004, Massart and Nédélec, 2006).

Assumption 1. *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 et al., 2016 (Papa et al., 2016) to show this.

Example 1. *Let N be a positive integer. For each vertex $i \in \{1, \ldots, n\}$, we observe $X_i = (X_i^1, X_i^2)$ where X_i^1 and X_i^2 are two distinct elements drawn from $\{1, \ldots, N\}$. 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* (1)*.*

B Universal Risk 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 3. *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

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

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 3, 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 common 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

$$\gamma : \mathcal{R} \times \mathcal{Z} \to [0, 1]. \tag{3}$$

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

$$\bar{\ell}(r^*, r) = P[\gamma(r, \cdot) - \gamma(r^*, \cdot)], \forall r \in \mathcal{R} \tag{4}$$

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

$$\gamma_{\mathbf{w}}(r) = \frac{1}{\|\mathbf{w}\|_1} \sum_{i=1}^{n} w_i \gamma(r, \xi_i) \tag{5}$$

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

$$\overline{\gamma_{\mathbf{w}}}(r) = \gamma_{\mathbf{w}}(r) - P[\gamma(r, \cdot)]. \tag{6}$$

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

$$\text{Var}[\gamma(r, \cdot) - \gamma(r^*, \cdot)] \leq d^2(r, r^*), \forall r \in \mathcal{R}. \tag{7}$$

A tighter risk bound for weighted ERM is derived from Theorem 3 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 3. *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 2. *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 4 (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 separability Assumption 2. Assume that, on the one hand,*

$$d(r^*,r) \leq \frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi(\sqrt{\bar{\ell}(r^*,r)}), \forall r \in \mathcal{R}, \tag{8}$$

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

$$\sqrt{\|\mathbf{w}\|_1}\mathbb{E}\left[\sup_{r' \in R', \frac{\|\mathbf{w}\|_2}{\sqrt{\|\mathbf{w}\|_1}}d(r',r)\leq\sigma} [\overline{\gamma_{\mathbf{w}}}(r') - \overline{\gamma_{\mathbf{w}}}(r)\right] \leq \phi(\sigma) \tag{9}$$

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

$$\sqrt{\|\mathbf{w}\|_1}\epsilon_*^2 = \phi(\psi(\epsilon_*)). \tag{10}$$

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

$$P\left[\bar{\ell}(r^*,\hat{r}) > 2\rho + 2\bar{\ell}(r^*,R) + Ky\epsilon_*^2\right] \leq e^{-y}. \tag{11}$$

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 1. *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 et al., 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 2 (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:*

$$\begin{aligned}\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j}f_{i,j}(X_i,X_j)|^q] \\ \leq 4^q\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j}f_{i,j}(X_i,X_j')|^q]\end{aligned} \tag{12}$$

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,

$$\begin{aligned}\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in\mathbb{E}} w_{i,j}f_{i,j}(X_i,X_j')|^q] \\ \leq 4^q\mathbb{E}[\sup_{f_{i,j}\in\mathcal{F}} |\sum_{(i,j)\in E} w_{i,j}f_{i,j}(X_i,X_j)|^q]\end{aligned} \tag{13}$$

Lemma 3 (randomization)**.** *Let $(\sigma_i)_{i=1}^n$ and $(\sigma_i')_{i=1}^n$ be two independent sequences of i.i.d. Radermacher 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 4. *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 $q \geq 1$,*

$$\begin{aligned}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i,X_j,Y_{i,j})|^q] \\ \leq 4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i,X_j',\tilde{Y}_{i,j})|^q]\end{aligned} \tag{14}$$

and the inequality can be reversed,

$$\begin{aligned}\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in\mathbb{E}} w_{i,j}\tilde{h}_r(X_i,X_j',\tilde{Y}_{i,j})|^q] \\ \leq 4^q\mathbb{E}[\sup_{r\in R}|\sum_{(i,j)\in E} w_{i,j}\tilde{h}_r(X_i,X_j,Y_{i,j})|^q]\end{aligned} \tag{15}$$

Lemma 5. *Let $(\sigma_i)_{i=1}^n$ and $(\sigma_i')_{i=1}^n$ be two independent sequences of i.i.d. Radermacher variables, independent from the $(X_i, X_i', Y_{i,j}, \tilde{Y}_{i,j})$'s. Then, we have for all $q \geq 1$,*

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

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

Lemma 2 and 3 are applications of Theorem 3.1.1 and Theorem 3.5.3 of (De la Pena and Giné, 2012) respectively, providing decoupling and randomization inequalities for degenerated weighted U-statistics of order 2. Lemma 4 and 5 are the variants of Lemma 2 and 3 respectively, which is suitable for the degenerated part $\widetilde{U}_{\mathbf{w}}(r)$.

Theorem 5 (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

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

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

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

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

$$P[Z \geq C\mathbb{E}[Z_\sigma] + t] \leq \exp\bigg(-\frac{1}{C}\min\Big((\frac{t}{\mathbb{E}[U_\sigma]})^2, \frac{t}{\mathbb{E}[M]+F\|\mathbf{w}\|_2}, (\frac{t}{\|\mathbf{w}\|_{\max}F})^{2/3}, \sqrt{\frac{t}{\|\mathbf{w}\|_\infty F}}\Big)\bigg)$$

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 $\mathscr{L}_\infty(\mathcal{X}^2)$ (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 E). Following the metric entropy inequality for Khinchine-type processes (see Section E), it is easy to get the following Corollary.

Corollary 1. *With the same setting of Theorem 5, if $\mathcal{F} \subset \mathscr{L}_\infty(\mathcal{X}^2)$, we have for any $\delta < 1/e$,*

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

where

$$\kappa \leq C\bigg(\|\mathbf{w}\|_2\int_0^{2F}\log N_\infty(\mathcal{F},\epsilon)d\epsilon + \max\Big(\|\mathbf{w}\|_2\log(1/\delta)\int_0^{2F}\sqrt{\log N_\infty(\mathcal{F},\epsilon)}d\epsilon, (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max}, (\log(1/\delta))^2\|\mathbf{w}\|_\infty\Big)\bigg).$$

with a universal constant C.

D Technical Proofs

D.1 Proofs Omitted in Section ??

Proof of Lemma **??**. Since $R \subset \mathscr{L}_\infty(\mathcal{X}^2)$ (Assumption **??**), 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

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

where

$$\kappa \leq \frac{C_1}{\|\mathbf{w}\|_1}\bigg(\|\mathbf{w}\|_2\int_0^1\log N_\infty(R,\epsilon)d\epsilon + \max\Big(\|\mathbf{w}\|_2\log(1/\delta)\int_0^1\sqrt{\log N_\infty(R,\epsilon)}d\epsilon, (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max}, (\log(1/\delta))^2\|\mathbf{w}\|_\infty\Big)\bigg).$$

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 **??**. Similarly, by Lemma 4 and Lemma 5, 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 5. More specifically, using (32) and (33) combined with the arguments in Corollary 1 and Assumption **??** will gives the second inequality for $\sup_{r\in R}|\widetilde{U}_{\mathbf{w}}(r)|$. □

Proof of Lemma **??**. For any function $r \in \mathcal{R}$, observe first that

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

Then observing that

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

almost sure, and combining with Jensen inequality, we have

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

□

Proof of Lemma **??**. First, we introduce some notations of weighted ERM of the i.i.d. case. We denote $\{\bar{w}_i = \sum_{j:(i,j)\in E} w_{i,j} : i = 1, \ldots, n\}$ 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

$$\phi(\sigma) = 12 \int_0^{\sigma} (\log N_\infty(R, \epsilon^2))^{1/2} d\epsilon. \tag{16}$$

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 **??**, 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

$$\psi(\epsilon) = \sqrt{5}\epsilon. \tag{17}$$

Then from Lemma 1, 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 4. For any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$\begin{aligned} L(r_{\mathbf{w}}) - L^* \leq 2(&\inf_{r \in R} L(r) - L^*) + 2\rho \\ &+ K \log(1/\delta)\epsilon_*^2) \end{aligned} \tag{18}$$

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 **??**, 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.2 Proof Omitted in Section A

Proof of Theorem 2. We write, for all $r \in R$,

$$\begin{aligned} L(r) - L_{\mathbf{w}}(r) &\leq \sup_{r \in R} \left[\mathbb{E}[\ell(r, z)] - \frac{1}{\|\mathbf{w}\|_1} \sum_{i=1}^{m} w_i \ell(r, z_i)\right] \\ &\leq \sum_{j=1}^{J} p_j \left[\sup_{r \in R} \sum_{i \in \mathcal{C}_j} \frac{w_{k_j^i}}{\|\mathbf{w}\|_1} (\mathbb{E}[\ell(r, z)] - \ell(r, z_{k_j^i}))\right]. \end{aligned} \tag{19}$$

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 $\mathbb{S}$, $f(\mathbb{S}) = \sum_{j=1}^{J} p_j f_j(S_{\mathcal{C}_j})$, then f satisfies the conditions of Theorem 1 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 $1 - \delta$,

$$\begin{aligned} L(r) - L_{\mathbf{w}}(r) \leq \mathbb{E}&\left[\sup_{r \in R}\left[\mathbb{E}[\ell(r, z)] - \frac{1}{\|\mathbf{w}\|_1} \sum_{i=1}^{m} w_i \ell(r, z_i)\right]\right] \\ &+ \sqrt{\chi^*(D_G) \log(1/\delta)} \frac{\|\mathbf{w}\|_2}{\|\mathbf{w}\|_1} \end{aligned}$$

Then, using the standard symmetrization technique (see Usunier et al., 2006 Usunier et al., 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. □

D.3 Proofs Omitted in Section B

Proof of Theorem 3. We use an 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

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

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. □

Proof of Theorem 4. Since R satisfies Condition 2, we notice that, by dominated convergence, for every $r \in R$, considering the sequence $\{r_k\}$ provided by Condition 2, 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

$$\bar{\ell}(r^*, \pi(r^*)) \leq \bar{\ell}(r^*, R) + \epsilon_*^2. \tag{20}$$

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

$$\begin{aligned}P[\bar{\ell}(r^*, \hat{r}) \leq 2(\rho + \bar{\ell}(r^*, \pi(r^*))) + 3\epsilon_*^2 + x^2] \\ \leq P[V_x \geq \frac{1}{2}].\end{aligned} \tag{21}$$

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 3. In order to use Theorem 3, we first remark that, by Condition 2,

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

which means that we indeed have to deal with a countably indexed empirical process. Note that the triangle inequality implies via (7), (20) and (8) that

$$\begin{aligned}(\mathrm{Var}[\gamma(r,\cdot) - \gamma(\pi(r^*),\cdot)])^{\frac{1}{2}} &\leq d(r^*, r) + d(r^*, \pi(r^*)) \\ &\leq 2\frac{\sqrt{\|\mathbf{w}\|_1}}{\|\mathbf{w}\|_2}\psi(\sqrt{\bar{\ell}(r^*, r) + \epsilon_*^2})\end{aligned} \tag{22}$$

Since γ takes its values in $[0, 1]$, introducing the functions $\psi_1 = \min(1, 2\psi)$ and, we derive from (22) that

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

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 3 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

$$V_x < \mathbb{E}[V_x] + \sqrt{\frac{2y(\psi_1^2(x)x^{-2} + 4\mathbb{E}[V_x])}{\|\mathbf{w}\|_1 x^2}} + \frac{2y}{3\|\mathbf{w}\|_1 x^2}. \tag{23}$$

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, (9) 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 (20), $\bar{\ell}(r^*, r) \leq a^2(r) \leq \bar{\ell}(r^*, r) + \epsilon_*^2$. 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 (20) 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 (8) 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 (9) 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

$$\mathbb{E}[V_x] \leq \frac{8\sqrt{2}\phi(\psi(\epsilon_*))}{\sqrt{\|\mathbf{w}\|_1}x\epsilon_*} = \frac{8\sqrt{2}\epsilon_*}{x} \leq \frac{8\sqrt{2}}{\sqrt{K'}y} \leq \frac{8\sqrt{2}}{\sqrt{K'}}, \tag{24}$$

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 (10) 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 (24) into (23) 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, (21) yields

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

We get the required probability bound (4) by setting $K = K' + 3$. □

Proof of Lemma 1. 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

$$\log \#\{\Pi_j \mathcal{F}\} \leq H_j \tag{25}$$

and

$$\Pi_j f \leq f \text{ with } P(f - \Pi_j f) \leq \sigma_j^2, \forall f \in \mathcal{F}. \tag{26}$$

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

$$P(|f - \Pi_0 f|) = P[|f - f_0|] \leq \sigma_0^2 = \sigma \tag{27}$$

for every $f \in \mathcal{F}$. Furthermore, since we may always assume that the extremities 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,

$$\begin{aligned}&\mathbb{E}[\sup_{f\in\mathcal{F}}[-v_{\mathbf{w}}(f)]]\\&\leq\sum_{j=0}^{J-1}\mathbb{E}[\sup_{f\in\mathcal{F}}v_{\mathbf{w}}(\Pi_j f)-v_{\mathbf{w}}(\Pi_{j+1}f)]+\sigma_J^2.\end{aligned}\tag{28}$$

Now, it follows from (26) and (27) 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, (25) 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 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 et al., 2017 Wang et al., 2017, Lemma 16) that

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

because $w_i\leq 1,\forall i\in 1,\ldots,n$, and (28) becomes

$$\begin{aligned}&\sqrt{\|\mathbf{w}\|_1}\mathbb{E}[\sup_{f\in\mathcal{F}}-v_{\mathbf{w}}(f)]\\&\leq 2\sum_{j=1}^{J}[\sigma_j\sqrt{5H_j}+\frac{1}{3\sqrt{\|\mathbf{w}\|_1}}H_j]+4\sqrt{\|\mathbf{w}\|_1}\sigma_{J+1}^2.\end{aligned}\tag{29}$$

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 (29) 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. □

D.4 Proofs Omitted in Section C

Proof of Lemma 4. This Lemma is derived from Lemma 2, thus we can follow the similar arguments that can be found in (De la Pena and Giné, 2012).

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:

$$\begin{aligned}&\mathcal{L}(X_1,\ldots,X_n,X_1',\ldots,X_n')=\\&\qquad\mathcal{L}(Z_1,\ldots,Z_n,Z_1',\ldots,Z_n')\end{aligned}\tag{30}$$

and

$$\mathcal{L}(X_1,\ldots,X_n)=\mathcal{L}(Z_1,\ldots,Z_n).\tag{31}$$

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,

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

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)

$$\begin{aligned}\mathbb{E}[\tilde{h}_r(X_i,X_j',\tilde{Y}_{i,j})\mid\Sigma]&=0\\\mathbb{E}[\tilde{h}_r(X_i',X_j,\tilde{Y}_{i,j})\mid\Sigma]&=0\end{aligned}$$

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

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

derived from the facts that the supreme and $|x|^p(p\geq 1)$ are convex functions and the Jansen inequality. According to (30) 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

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

which concludes the proof of (14).

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

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

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

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

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

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

(According to (31))

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

which (15) follows. □

Proof of Lemma 5. Re-using the notations used in the proof of Lemma 2, 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:

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

($\tilde{h}_r$ is degenerated)

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

(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:

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

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

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

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 together. □

Proof of Theorem 5. By the decoupling, undecoupling and randomization techniques (see Lemma 2, Lemma 3), 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

$$\mathbb{E}[\mathbb{E}_\sigma[Z'_\sigma]^q] \leq 2^{q-1}(2^q(E[Z'_\sigma])^q + 5^q q^q\mathbb{E}[\mathbb{E}_\sigma[M]^q]) \quad (32)$$

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

$$\mathbb{E}[\mathbb{E}_\sigma[M]^q] \leq 2^{q-1}(2^q\mathbb{E}[M]^q + 5^q q^q\|\mathbf{w}\|_\infty^q F^q). \quad (33)$$

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-Schwartz 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-Schwartz 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

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

for an appropriate constant C. In order to derive the exponential inequality, we use Markov 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. □

Proof of Corollary 1. From Theorem 3, it is easy to know

$$\begin{aligned}\kappa =& C(\mathbb{E}[Z_\sigma] + \max(\mathbb{E}[U_\sigma]\sqrt{\log(1/\delta)}, \mathbb{E}[M]\log(1/\delta),\\ & \log(1/\delta)\|\mathbf{w}\|_2, (\log(1/\delta))^{3/2}\|\mathbf{w}\|_{\max},\\ & (\log(1/\delta))^2\|\mathbf{w}\|_\infty)).\end{aligned} \tag{34}$$

An important character of these Rademacher processes in (34) is that they all satisfy the Khinchine inequality (36). 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 6, we can easily have that z'_σ satisfies (36) with degree 2. Thus from Theorem 7, we have

$$\mathbb{E}_\sigma[\sup_{f,g}|z'_\sigma(f) - z'_\sigma(s)|] \leq K\int_0^D \log N(\mathcal{F}, \mathbb{L}_2, \epsilon)d\epsilon. \tag{35}$$

Recall that $\sup_f\|f\|_\infty = F$, we have $D = 2F$. Since this metric function d is intractable, we need to convert it to the $\mathbb{L}_\infty$ metric. For all f, s, we have $\mathbb{L}_2(z'_\sigma(f), z'_\sigma(g)) \leq \mathbb{L}_\infty(f,g)$. The fact that if $\forall f, s \in \mathcal{F}, \mathbb{L}(f,s) \leq \mathbb{L}'(f,s)$ then $N(\mathcal{F}, \mathbb{L}, \epsilon) \leq N(\mathcal{F}, \mathbb{L}', \epsilon)$ combined with (35) will give

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

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 Khintchine inequality with degree 1. Also, we need to convert its metric distance and $\mathbb{L}_2(u_\sigma(f), u_\sigma(g)) \leq \mathbb{L}_\infty(f,g)$. Thus, $\mathbb{E}[U_\sigma] \leq K\|\mathbf{w}\|_2\int_0^{2F}\sqrt{\log N_\infty(\mathcal{F}, \epsilon)}d\epsilon$.

Plugging all these part into (34) will complete the corollary. □

E Metric Entropy Inequality

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

Theorem 6 (Khinchine inequality for Rademacher chaos, De la Pena and Giné, 2012 (De la Pena and Giné, 2012, Theorem 3.2.1))**.** *Let F be a normed vector space and let $\{\sigma_i\}_{i=1}^\infty$ be a Rademacher sequence. Denote by*

$$\begin{aligned}X = x + \sum_{i=1}^n x_i\sigma_i + \sum_{i_1<i_2\leq n} x_{i_1,i_2}\sigma_{i_1}\sigma_{i_2} + \dots\\ + \sum_{i_1<\dots<i_d\leq n} x_{i_1\dots i_d}\sigma_{i_1}\dots\sigma_{i_d}\end{aligned}$$

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 $d \geq 1$,

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

Theorem 7 (metric entropy inequality, Arcones and Gine, 1993 (Arcones and Gine, 1993, Proposition 2.6))**.** *If a process $\{Y_f : f \in \mathcal{F}\}$ satisfies*

$$(\mathbb{E}[|Y_f - Y_g|^p])^{1/p} \leq (\frac{p-1}{q-1})^{m/2}(\mathbb{E}[|Y_f - Y_g|^q])^{1/q}, \tag{36}$$

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

$$d(f,g) = (\mathbb{E}[|Y_f - Y_g|^2])^{1/2}, \tag{37}$$

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

$$\mathbb{E}[\sup_{f,g}|Y_f - Y_g|] \leq K\int_0^D (\log N(\mathcal{F}, d, \epsilon))^{m/2}d\epsilon. \tag{38}$$

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

References

Arcones, M. A. and Gine, E. (1993). Limit theorems for u-processes. *The Annals of Probability*, pages 1494–1542.

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

Boucheron, S., Bousquet, O., Lugosi, G., Massart, P., et al. (2005). Moment inequalities for functions of independent random variables. *The Annals of Probability*, 33(2):514–560.

Bousquet, O. (2002). A bennett concentration inequality and its application to suprema of empirical processes. *Comptes Rendus Mathematique*, 334(6):495–500.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). Ranking and empirical minimization of u-statistics. *Annals of Statistics*, 36(2):844–874.

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

De la Pena, V. and Giné, E. (2012). *Decoupling: from dependence to independence*. Springer Science & Business Media.

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

Massart, P. and Nédélec, É. (2006). Risk bounds for statistical learning. *Annals of Statistics*, pages 2326–2366.

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*, pages 694–702.

Ralaivola, L. and Amini, M.-R. (2015). Entropy-based concentration inequalities for dependent variables. In *International Conference on Machine Learning*, pages 2436–2444.

Ralaivola, L., Szafranski, M., and Stempfel, G. (2009). Chromatic pac-bayes bounds for non-iid data. In *Artificial Intelligence and Statistics*, pages 416–423.

Tsybakov, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, pages 135–166.

Usunier, N., Amini, M.-R., and Gallinari, P. (2006). Generalization error bounds for classifiers trained with interdependent data. In *Advances in neural information processing systems*, pages 1369–1376.

Wang, Y., Guo, Z.-C., and Ramon, J. (2017). Learning from networked examples. In *Proceedings of the 28th International Conference on Algorithmic Learning Theory*, page to appear.