HAL open science

Learning from Networked Examples

Yuyi Wang, Zheng-Chu Guo, Jan Ramon

To cite this version:

Yuyi Wang, Zheng-Chu Guo, Jan Ramon. Learning from Networked Examples. ALT 2017 - 28th conference on Algorithmic Learning Theory, Oct 2017, Kyoto, Japan. pp.1 - 26. hal-02519274

HAL Id: hal-02519274

https://inria.hal.science/hal-02519274v1

Submitted on 25 Mar 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Proceedings of Machine Learning Research 76:1–26, 2017 Algorithmic Learning Theory 2017

Learning from Networked Examples

Yuyi Wang YUWANG@ETHZ.CH
ETH Zürich, Switzerland

Zheng-Chu Guo GUOZHENGCHU@ZJU.EDU.CN
Zhejiang University, P.R. China

Jan Ramon JAN.RAMON@CS.KULEUVEN.BE
Inria, France & KULeuven, Belgium

Editors: Steve Hanneke and Lev Reyzin

Abstract

Many machine learning algorithms are based on the assumption that training examples are drawn independently. However, this assumption does not hold anymore when learning from a networked sample because two or more training examples may share some common objects, and hence share the features of these shared objects. We show that the classic approach of ignoring this problem potentially can have a harmful effect on the accuracy of statistics, and then consider alternatives. One of these is to only use independent examples, discarding other information. However, this is clearly suboptimal. We analyze sample error bounds in this networked setting, providing significantly improved results. An important component of our approach is formed by efficient sample weighting schemes, which leads to novel concentration inequalities.

Keywords: Networked Example, Non-i.i.d. Sample, Concentration Inequality, Sample Error Bound, Generalization Error Bound

1. Introduction

Recently, there has been an increasing interest in network-structured data, such as data in social networks (Leinhardt, 2013), economic networks (Knoke, 2014), citation networks (Nykl et al., 2014) and chemical interaction networks (Szklarczyk et al., 2014). An important challenge is that data concerning related objects cannot be assumed to be independent. More precisely, a key assumption made by many approaches in the field of statistics and machine learning is that observations are drawn independently and identically (i.i.d.) from a fixed distribution. However, this assumption does not hold anymore for observations extracted from a network.

Many practical approaches to supervised learning in networks ignore (at least partially) the problem and learn models with classic machine learning techniques. While these work to some extent, they are not supported by a well-developed theory such as the one which provides generalization guarantees for the i.i.d. case as a function of the number of training examples. As a consequence, one may miss opportunities to learn due to the numerous dependencies between the training examples.

In this paper, we make a step towards addressing this problem. More specifically, the contributions of this paper are as follows:

© 2017 Y. Wang, Z.-C. Guo & J. Ramon.

- We introduce a framework for explicitly representing dependencies among examples. Such framework allows for encoding domain knowledge and can form a basis for studies preparing a machine learning effort on networked data.
- We introduce a relaxed i.i.d. assumption for networked examples, which generalizes over classic i.i.d. assumptions and covers a number of existing learning problems.
- We show new concentration inequalities for networked examples and demonstrate the applicability of our results to learning theory by upgrading an empirical risk minimization result to networked data. In the process, we improve on earlier concentration inequalities by Janson (2004) and then improve on earlier learning guarantees, even if we use the original unweighted ERM.
- We propose a novel learning scheme based on appropriate weighting on networked examples, which is efficiently implementable by solving a linear program and satisfies monotonicity which ensures that one can always do better when the networked dataset becomes larger (in the sense of super-hypergraphs).

1.1. Related work

Usunier et al. (2006) define interdependent training data which is closely related to networked examples, make similar assumptions, and show generalization error bounds for classifiers trained with this type of data. In (Ralaivola et al., 2010), PAC-Bayes bounds for classification with non-i.i.d data are established based on fractional colorings of graphs, results also hold for specific learning settings such as ranking and learning from stationary β- mixing distributions. (Ralaivola and Amini, 2015) establishes new concentration inequalities based on the idea of fractional chromatic numbers and entropy methods for fractionally sub-additive and fractionally self-bounding functions of dependent variables. All these results are based on Janson's work (Janson, 2004) and use dependency graphs (Erdos and Lovász, 1975) to represent the examples and their relations. In this paper, we use a different representation for networked examples, improve the concentration inequalities in (Janson, 2004) and then potentially improve on these existing learning results. Besides, we propose a better learning scheme than that in these existing works.

Wang et al. (2011) considered a similar setting of networked examples. Their work assumes a bounded covariance between pairs of examples connected with an edge (excluding possible higher-order interactions) but this is not assumed in our model. While we use our model to show learning guarantees, Wang et al. (2011) show corrections for the bias (induced by the dependencies between examples) on statistical hypothesis tests.

Mixing conditions are also used to model non-i.i.d. samples. For example, in (Sun and Wu, 2010; Guo and Shi, 2011), the learning performance of regularized classification and regression algorithm using a non-i.i.d. sample is investigated, with the assumption that the training sample satisfies mixing conditions. Modha and Masry (1996) presented a Bernstein-type inequality for stationary exponentially α-mixing processes based on the effective number (less than the sample size). Our Bernstein-type inequalities for networked random variables assigns weights to examples. However, the assumptions for the training sample are different, and the main techniques are distinct. We refer interested readers to (Bradley, 2005) and references therein for more details about the mixing conditions.

Our theory is applicable to problems in the field of Statistical Relational Learning (SRL) (Getoor and Taskar, 2007), e.g., for learning local conditional probability functions for directed models such as Probabilistic Relational Models (Friedman et al., 1999), Logical Bayesian Networks (Fierens et al., 2005), Relational Bayesian Networks (Jaeger, 1997). There is a huge literature in SRL for learning features and existence of edges in a graph, for which we refer the reader to the survey by Rossi et al. (2012).

2. Model

In this section, we introduce networked examples. The basic intuition is that a networked example combines a number of objects, each of which can carry some information (features). Every object may be shared among several examples, such that we get a network. The sharing of information also makes explicit the dependencies between the examples.

2.1. Networked examples

We use a hypergraph $G = (V_G, E_G)$ to represent a set of networked examples. The vertices $V_G = \{v_1, v_2, \ldots, v_m\}$ represent objects, and the hyperedges $E_G = \{e_1, e_2, \ldots, e_n\}$ represent examples grouping a number of objects. In what follows, we will often abuse terminology, identify vertices and objects and use the terms interchangeably.

We denote with $\mathcal{X}$ the space of features of objects. These features can be represented with feature vectors or other suitable data structures. As usual in supervised learning, an example is a pair of an input and a target value. The input of an example is a collection of features of objects and is represented by a multiset of elements of $\mathcal{X}$. We denote with $\mathbb{X}$ the space of features of all possible examples. Each training example has a target value, e.g., a class label or regression value. We denote with $\mathcal{Y}$ the space of all possible target values.

This hypergraph-based representation together with the weaker independence assumption (see the next subsection) have powerful expressive ability and cover many existing models. For example, many previous works treat the ranking problem as binary classification on networked data.

Learning to rank *Consider the pairwise document ranking problem, which has been extensively studied for information retrieval (Liu et al., 2009). We have a hypergraph G with a vertex set V_G of documents, and every hyperedge contains two document vertices (this hypergraph is actually a normal graph since every hyperedge has exactly two vertices). Every document is described by a feature vector from $\mathcal{X}$ which could be a bag of words; every training example contains two documents (and their features from $\mathbb{X} = \mathcal{X} \times \mathcal{X}$). The target value of an example is the order between the two documents, i.e., $\mathcal{Y} = \{0, 1\}$ and 0 means that the first document is ranked higher than the second one, 1 otherwise. In fact, an even larger class of learning to rank problems can be modeled using networked examples (called* interdependent examples *in the literature), and we refer the interested readers to (Amini and Usunier, 2015, Chap. 4).* ♦

In the following example, we make a simplified version of the problem of applying machine learning techniques to decide which projects to invest. In contrast to the example above, every hyperedge in the example of investment may have more than two vertices and the cardinality of every hyperedge may be different from each other.

Investment *Some organizations such as banks or research councils provide financial-related support, e.g., loans or grants, to business projects or research projects. These organizations may use*

machine learning techniques to help deciding which projects to invest. In the hypergraph-structured training dataset, every vertex is a person with a feature vector from $\mathcal{X}$ describing education level, age, ... and every training example contains a group of people who take charge of a project. The target value of an example is the return on investment. It is common that some people may be involved in several different projects, and in this way we have a hypergraph. ♦

Our model and results can also be applied to an important subclass of hypergraphs in which V_G is a union of disjoint sets $\{V_G^{(i)}\}_{i=1}^k$ and $E_G \subseteq \times_{i=1}^k V_G^{(i)}$, i.e., k-partite hypergraphs.

Movie rating *Consider the problem of movie rating. We have a tripartite hypergraph G with a vertex set $V_G^{(1)}$ of persons, a vertex set $V_G^{(2)}$ of movies and a vertex set $V_G^{(3)}$ of cinemas. Every hyperedge contains a person vertex, a movie vertex and a cinema vertex. The feature space $\mathcal{X}$ also has three parts $\mathcal{X}^{(1)}$, $\mathcal{X}^{(2)}$ and $\mathcal{X}^{(3)}$. Every person (a vertex in $V_G^{(1)}$) can be described by a feature vector from $\mathcal{X}^{(1)}$ with gender, age, nationality, ... Every movie (a vertex in $V^{(2)}$) can be described by a feature vector from $\mathcal{X}^{(2)}$ with genre, actor popularity, ... Every cinema vertex (a vertex in $V^{(3)}$) can be described by a feature vector from $\mathcal{X}^{(3)}$ with location, equipments, ... Then, $\mathbb{X} = \mathcal{X}^{(1)} \times \mathcal{X}^{(2)} \times \mathcal{X}^{(3)}$ is the space of feature vectors of complete examples, consisting of a concatenation of a person feature vector, a movie feature vector and a cinema feature vector. The target value of an example is the rating the person gave to the movie in the concerned cinema, e.g., the space $\mathcal{Y}$ can be the set $\{1, 2, \ldots, 10\}$. A trained model may help recommending movies and suggest good cinemas to watch the movie.* ♦

2.2. Independence assumption

Though networked examples are not independent, we still need to assume some weaker form of independence of the examples. If we would not make any assumption, the dependence between examples could be so strong that they perfectly correlate. In such situation, it is not possible to learn.

We consider a labeled hypergraph $(G, \mathcal{X}, \phi, \mathcal{Y}, \lambda)$, where the labels assigned by $\phi : V_G \mapsto \mathcal{X}$ and $\lambda : E_G \mapsto \mathcal{Y}$ are drawn randomly from a probability distribution ρ. We make the following assumptions:

- Features of every object (assigned to vertices) are independent from features of other objects, i.e., there is a probability distribution $\rho_{\mathcal{X}}$ on $\mathcal{X}$ such that for every $q \in \mathcal{X}$ and $v \in V_G$, $\rho(\phi(v) = q \mid \phi(u)) = \rho(\phi(v) = q) = \rho_{\mathcal{X}}(q)$ for any $u \neq v$. It is not required that objects have an identical distribution, though we use the same notation $\rho_{\mathcal{X}}$ for all objects for the sake of simplicity.

- Moreover, every example (assigned to hyperedges) gets a target value drawn independently given the features of the objects (vertices) incident with the hyperedge, i.e., there is a probability distribution $\rho_{\mathcal{Y}|\mathbb{X}}$ on $\mathcal{Y} \times \mathbb{X}$ such that for all $e \in E_G$, $\rho(\lambda(e) = y \mid \phi|_e) = \rho_{\mathcal{Y}|\mathbb{X}}(y, \phi|_e) = \rho(\lambda(e) = y \mid \phi, E_G)$. Here, $\phi|_e$ is ϕ restricted to e, i.e., $\phi|_e = \{(v, \phi(v)) \mid v \in e\}$.

From the above assumptions, we can infer that

$$\rho(\phi, \lambda) = \prod_{v \in V_G} \rho_{\mathcal{X}}(\phi(v)) \prod_{e \in E_G} \rho_{\mathcal{Y}|\mathbb{X}}(\lambda(e), \phi|_e).$$

Our analysis holds no matter what the distribution ρ is, as long as the above assumptions are met.

It is worth pointing out that this weak independence assumption may not yet hold in all real-world situations but is already a better approximation than the classic i.i.d. assumptions. For instance, consider again the problem of learning to rank, this weak independence assumption is usually satisfied, and whether we know the target value of a specific pair of documents only depends on whether it is in the collected data, independent of the features of the two documents. Besides, similar assumptions are also made in several previous works such as (Usunier et al., 2006; Ralaivola et al., 2010; Ralaivola and Amini, 2015).

3. Effective sample size, intuition, and examples

An important aspect of the theory presented in this paper can be understood as a better estimation of the *effective sample size* of networked datasets. Generally speaking one can define the effective sample size of a dataset G, for a particular statistical task of estimating T, as the number of examples an i.i.d. dataset would need to allow for estimating T as accurately as can be done with the networked dataset G. In this paper, we consider *mean value estimation* and use concentration inequalities as references.

Classical concentration inequalities such as Bernstein inequality (Bernstein, 1924), Bennett inequality (Bennett, 1962) and Chernoff-Hoeffding inequalities (Chernoff, 1952; Hoeffding, 1963) (see e.g., (Chung and Lu, 2006, Chap. 2)) are used to analyze generalization errors of learning problems on i.i.d. samples.

Janson (2004) showed concentration inequalities (see Appendix A) for the (equally weighted) mean value estimator of *networked random variables* by the so-called fractional-coloring-based method.

Definition 1 (Networked random variables) *Given a hypergraph G, we call $\{\xi_i\}_{i=1}^n$ G-*networked random variables *if there exist functions $f_i : \mathbb{X} \mapsto \mathbb{R}$ such that $\xi_i = f_i(\{\Phi_v \mid v \in e_i\})$ where $\{\Phi_v\}_{v\in V}$ is a set of independent random variables indexed by the vertices of G and $\{\Phi_v \mid v \in e_i\} = x_i$ is the feature of example i.*

In Janson's inequalities, a key parameter is the fractional hyperedge-chromatic number[1] χ_G^*.

Definition 2 (Fractional hyperedge-chromatic number) *A b-fold hyperedge-coloring of a hypergraph G is an assignment of b colors to every hyperedge in E_G such that adjacent hyperedges have no color in common. The b-fold hyperedge-chromatic number $\chi_G^{(b)}$ of a hypergraph G is the smallest number of colors needed to obtain a b-fold hyperedge-coloring of the hyperedges in E_G. The fractional hyperedge-chromatic number χ_G^* is*

$$\chi_G^* = \lim_{b\to\infty} \frac{\chi_G^{(b)}}{b} = \inf_b \frac{\chi_G^{(b)}}{b}.$$

Though Janson's result and its variations are the foundation of (almost) all existing theoretical work related to learning from networked examples (Usunier et al., 2006; Ralaivola et al., 2010; Amini and Usunier, 2015), we begin our technical discussion with the following examples which show that it is necessary and possible to improve Janson's result (also see the remark after Theorem 14). We denote the effective sample size with Janson's method by CESS, i.e., $\text{CESS}(G) = n/\chi_G^*$,

1. In (Janson, 2004), this parameter is called the fractional cover number.

and denote the effective sample size with our new method by MESS. According to the results in Sec. 4, $\text{MESS}(G) = \nu_G^*$ the fractional matching number of G (in Example 1 and 2, $\text{MESS}(G) = \nu_G^* = n/\omega_G$ where ω_G is the maximum degree of G).

Example 1 (MESS $= \Omega(n) \cdot$ CESS) *Consider two hypergraphs each of which has n hyperedges:*

- *G_n: all the n hyperedges share a common vertex; and*
- *H_n: all the n hyperedges intersect with each other, but there do not exist any three hyperedges that share a common vertex.*

For any n, hypergraphs G_n and H_n have the same fractional hyperedge-chromatic number $\chi^*_{G_n} = \chi^*_{H_n} = n$ because in both cases any two hyperedges intersect with each other. According to Janson's result, both of them have effective sample size $n/\chi^*_{G_n} = n/\chi^*_{H_n} = 1$.

For hypergraph G_n, this result is reasonable because all the networked random variables may be highly controlled by the shared vertex. However, if two networked random variables in H_n are highly correlated, then intuitively the two variables cannot be very correlated with other networked random variables in H_n; hence the effective sample size of H_n would be larger than 1. This is indeed the case, as we show in Corollary 7 the effective sample size of H_n can be $n/2$ which cannot be obtained by fractional-coloring-based approaches because these methods only consider which examples are dependent, but not the details - how they are dependent. ***Details matter!***

The example above shows, that if there is a subset of networked random variables that any two of them overlap, then fractional-coloring-based approaches only deal with the worst case that all the networked random variables share a common vertex but ignore the details of the hypergraph. A natural question to ask is whether this is the "only way" that MESS can beat CESS. In particular, are there other cases without "the issue of details" that MESS can still be larger than CESS? The following example shows that the answer is yes.

Example 2 (MESS $>$ CESS on odd cycles) *Consider an odd cycle G with $n = 2k + 1$ hyperedges, i.e., hyperedge e_i only intersect with two hyperedges $e_{(i-1) \mod n}$ and $e_{(i+1) \mod n}$. The fractional hyperedge-chromatic number of G_n is $2+1/k$ (see e.g.,(Scheinerman and Ullman, 2011)), so $\text{CESS}(G_n) = n/(2 + 1/k) = k$. By Corollary 7, we improve the result to $\text{MESS}(G_n) = n/2 = k + 1/2$. Note that in this example, there does not exist the issue of details mentioned in Example 1, but a gap between* MESS *and* CESS *still exists.*

In both examples above all the hypergraphs we considered have *symmetric* structures, the equally weighted estimator seems work well. But, the following example shows that this simple estimator is not always good, as it does not satisfy *monotonicity*: Roughly speaking, given n i.i.d. examples, when the sample size n is larger then we would expect a more accurate estimator. Similarly, if G' is a super-hypergraph of G, then the effective sample size of G'-networked examples should be not smaller than that of G-networked examples. The intuition is that if we have more data then we should learn better (at least not worse), or else we just select a subsample to learn.

Example 3 (Monotonicity) *Consider the following hypergraphs:*

- *G: has $n_1 + n_2$ hyperedges in total, in which n_1 hyperedges are independent and the rest n_2 hyperedges share a common vertex.*

- *H: is a super-hypergraph of G, and it has $n_1 + n_2 + 1$ hyperedges in total, in which n_1 hyperedges are independent and the rest $n_2 + 1$ hyperedges share a common vertex.*

The fractional hyperedge-chromatic number of G and H are n_2 and $n_2 + 1$ respectively (assume $n_1, n_2 \geq 1$). Therefore, $\text{CESS}(G) = (n_1+n_2)/n_2$ is greater than $\text{CESS}(H) = (n_1+n_2+1)/(n_2+1)$ if we use this simple estimator, though H is a super-hypergraph of G.

In the example above, why don't we just ignore one of the $(n_2 + 1)$ hyperedges in H and then obtain a better effective sample size that is the same as $\text{CESS}(G)$? This idea can be generalized: We just select a *maximum matching* from the hypergraph, and then only use the examples corresponding to this matching. It is easy to see that

- this method satisfies the monotonicity since the matching number ν of a hypergraph is always not larger than that of its super-hypergraph;
- besides, it is well known that $\nu_G \geq n/\chi^*_G$ for all hypergraphs, so it seems that the matching-based method can always achieve larger effective sample size than the fractional-coloring-based method.

However, the matching-based method is still a suboptimal solution, because

- it is NP-hard to compute the maximum matching of a hypergraph in general,
- though $\nu_G \geq n/\chi^*_G$ for all hypergraphs, we show out that n/χ^*_G is not tight for the equally weighted estimator, so matching-based method is not always better than the equally weighted estimator.

In order to tackle this problem and accurately estimate the mean value[2] $\mu = \mathbb{E}[\xi]$ using networked examples, we consider a class of estimators of the form

$$\hat{\mu}_{f,w} = \frac{\sum_{i=1}^n w_i \xi_i}{\sum_{i=1}^n w_i}$$

and three weighting schemes

- EQW: all examples get equal weights, i.e., $w_i = 1$ for all i. As far as we know, almost all related works only deal with estimators of this type.
- IND: a maximum-size set $E_{IND} \subseteq E$ of independent examples is selected, i.e., $\forall e_1, e_2 \in E_{IND} : e_1 \cap e_2 = \emptyset$. Examples in E_{IND} are weighted 1, otherwise weighted 0.
- FMN: weight the examples with a fractional matching; it is introduced in Section 4.1.

4. New concentration inequalities

In this section, we show new concentration bounds if random variables are allowed to have non-identical weights. These concentration inequalities cannot be obtained by Janson's method (the fractional-coloring-based method), so we prove new lemmas. In the next section, we use these results to bound generalization errors for learning from networked examples.

2. For the sake of simplicity, we assume every ξ_i has the same expected value, and our analysis can be generalized to the case that ξ_i's have non-identical expected values.

4.1. Fractional matching schemes

Before stating the main result (Theorem 6), we first define fractional matchings.

Definition 3 (Fractional matching) *Given a hypergraph $G = (V_G, E_G)$ with $E_G = \{e_i\}_{i=1}^n$, a fractional matching w is a nonnegative vector $(w_i)_{i=1}^n$ defined on the hyperedges satisfying that for every vertex $v \in V_G$, $\sum_{i:v\in e_i} w_i \leq 1$. In other words, a weight vector is a fractional matching if for every vertex the sum of the weights of the incident hyperedges is at most 1.*

Remark: The IND weighting scheme is also a (fractional) matching method and every weight w_i must be either 0 or 1, i.e., these vectors correspond to matchings in hypergraphs. The original concentration inequalities can be applied directly to the IND weighting scheme, since all the examples with weight 1 are mutually independent. Besides, since larger matchings provide tighter bounds, we tend to find a maximum matching whose sum is called the *matching number* and is denoted by ν_G. However, it is in general an NP-hard problem to find a maximum matching in hypergraphs (see e.g., (Garey and Johnson, 1979)). Moreover, the maximum matching problem is also an APX-complete problem (Uriel et al., 1991), so we would not expect an efficient algorithm to achieve a good approximation either.

A key property used for proving classical concentration inequalities is that all observations are independent. That is, if $\{\xi_i\}_{i=1}^n$ are independent random variables, then the moment-generating function $\mathbb{E}[\exp(c\sum_{i=1}^n \xi_i)]$, where $c \in \mathbb{R}$, satisfies

$$\mathbb{E}\left[\exp\left(c\sum_{i=1}^n \xi_i\right)\right] = \prod_{i=1}^n \mathbb{E}\left[e^{c\xi_i}\right].$$

However, when considering networked random variables, the equality does not hold anymore. Instead, we show a new property in Lemma 4.

Lemma 4 *Given G-networked random variables $\{\xi_i\}_{i=1}^n$, if $w = (w_i)_{i=1}^n$ is a fractional matching of the hypergraph G, then*

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^n w_i\xi_i\right)\right] \leq \prod_{i=1}^n \left(\mathbb{E}\left[e^{\xi_i}\right]\right)^{w_i}. \tag{1}$$

Proof First, note that the expectation in the left hand side of Inequality (1) is over the (independent) features $S_1, \ldots, S_{|V_G|}$ of the vertices of G, because these are the basic random variables of which the $(\xi_i)_{i=1}^n$ are composed. We prove this theorem by induction on $|V_G|$. For $|V_G| = 1$,

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^n w_i\xi_i\right)\right] = \mathbb{E}_{S_1}\left[\prod_{i=1}^n e^{w_i\xi_i}\right].$$

Using Lemma 15 with $t = \left(e^{\xi_i}\right)_{i=1}^n$, $\beta = w$ and $g(t) = \prod_{i=1}^n e^{w_i\xi_i}$, we know that $g(t)$ is a concave function since w is a fractional matching. Given that $g(t)$ is concave, we have

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^n w_i\xi_i\right)\right] = \mathbb{E}_{S_1}[g(t)] \leq g(\mathbb{E}_{S_1}[t]) = \prod_{i=1}^n \left(\mathbb{E}\left[e^{\xi_i}\right]\right)^{w_i}$$

which follows from Jensen's inequality (Jensen, 1906). Assume that the theorem is true for $|V_G| = 1, \ldots, m-1$, we now prove the theorem for $|V_G| = m$. We can write

$$\mathbb{E}\left[\exp\left(\sum_{i=1}^{n} w_i \xi_i\right)\right] = \mathbb{E}_{S_m}\left[\mathbb{E}\left[\prod_{i=1}^{n} e^{w_i \xi_i} \middle| S_m\right]\right]. \tag{2}$$

where the $\mathbb{E}[\cdot|\cdot]$ notation on the right hand side denotes a conditional expectation. We use the induction hypothesis on the right hand side of Eq. (2), yielding

$$\mathbb{E}_{S_m}\left[\mathbb{E}\left[\prod_{i=1}^{n} e^{w_i \xi_i} \middle| S_m\right]\right] \leq \mathbb{E}_{S_m}\left[\prod_{i=1}^{n} \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)^{w_i}\right]. \tag{3}$$

We define two index sets A and B, partitioning hyperedges in G (and hence random variables ξ_i) into a part which is incident with v_m (dependent on S_m) and a part which is not, i.e., $A := \{i | v_m \in e_i\}$ and $B := \{i | v_m \notin e_i\}$. Then, for all $i \in B$, ξ_i is independent of S_m. We can write this as

$$\mathbb{E}_{S_m}\left[\prod_{i=1}^{n} \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)^{w_i}\right] = \mathbb{E}_{S_m}\left[\prod_{i \in A} \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)^{w_i}\right] \prod_{i \in B} \left(\mathbb{E}\left[e^{\xi_i}\right]\right)^{w_i}. \tag{4}$$

Let $t = \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)_{i \in A}$, $\beta = (w_i)_{i \in A}$ and $g(t) = \prod_{i \in A} \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)^{w_i}$. According to the definition of fractional matchings and Lemma 15, we know that $g(t)$ is concave. Again, by Jensen's inequality, we have

$$\mathbb{E}_{S_m}\left[\prod_{i \in A} \left(\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right)^{w_i}\right] \leq \prod_{i \in A} \left(\mathbb{E}_{S_m}\left[\mathbb{E}\left[e^{\xi_i} \middle| S_m\right]\right]\right)^{w_i} = \prod_{i \in A} \left(\mathbb{E}\left[e^{\xi_i}\right]\right)^{w_i}. \tag{5}$$

From Equations (2), (4) and Inequalities (3) and (5), we can see that this theorem holds for $|V_G| = m$. ■

Remark: Similar results hold for all nonnegative functions of ξ, not only e^{ξ}. More precisely, given G-networked random variables $\{\xi_i\}_{i=1}^n$, it holds that $\mathbb{E}[\prod_{i=1}^n (f(\xi_i))^{w_i}] \leq \prod_{i=1}^n (\mathbb{E}[f(\xi_i)])^{w_i}$ for any nonnegative function f and any fractional matching w of G.

The follow result, which is the networked analogues of the Bennett inequality, shows how to use Lemma 4 to prove new concentration inequalities.

Lemma 5 *Let $(\xi_i)_{i=1}^n$ be G-networked random variables with mean $\mathbb{E}[\xi_i] = \mu$ and variance $\sigma^2(\xi_i) = \sigma^2$, such that $|\xi_i - \mu| \leq M$ with probability 1. Let $w = (w_i)_{i=1}^n$ be a fractional matching for G, and let $|w| = \sum_i w_i$, then for all $\epsilon > 0$,*

$$\Pr\left(\sum_i w_i (\xi_i - \mu) \geq \epsilon\right) \leq \exp\left(-\frac{|w|\sigma^2}{M^2} h\left(\frac{M\epsilon}{|w|\sigma^2}\right)\right)$$

where $h(a) = (1+a)\log(1+a) - a$ for any real number a.

Proof Without loss of generality, we assume $\mu = 0$. Let c be an arbitrary positive constant which will be determined later. Then

$$I := \Pr\left(\sum_{i=1}^{n} w_i\xi_i \geq \epsilon\right) = \Pr\left(\exp\left(c\sum_{i=1}^{n} w_i\xi_i\right) \geq e^{c\epsilon}\right).$$

By Markov's inequality and Lemma 4, we have

$$I \leq e^{-c\epsilon}\mathbb{E}\left[\exp\left(c\sum_{i=1}^{n} w_i\xi_i\right)\right] \leq e^{-c\epsilon}\prod_i \left(\mathbb{E}\left[e^{c\xi_i}\right]\right)^{w_i}.$$

Since $|\xi_i| \leq M$ and $\mu = 0$, we have

$$\mathbb{E}\left[e^{c\xi_i}\right] = 1 + \sum_{p=2}^{+\infty}\frac{c^p\mathbb{E}[\xi_i^p]}{p!} \leq 1 + \sum_{p=2}^{+\infty}\frac{c^pM^{p-2}\sigma^2}{p!}$$

from the Taylor expansion for exponential functions. Using $1 + a \leq e^a$, it follows that

$$\mathbb{E}\left[e^{c\xi_i}\right] \leq \exp\left(\sum_{p=2}^{+\infty}\frac{c^pM^{p-2}\sigma^2}{p!}\right) = \exp\left(\frac{e^{cM} - 1 - cM}{M^2}\sigma^2\right)$$

and therefore

$$I \leq \exp\left(-c\epsilon + \frac{e^{cM} - 1 - cM}{M^2}|w|\sigma^2\right).$$

Now choose the constant c to be the minimizer of the bound on the right hand side above:

$$c = \frac{1}{M}\log\left(1 + \frac{M\epsilon}{|w|\sigma^2}\right).$$

That is, $e^{cM} - 1 = \frac{M\epsilon}{|w|\sigma^2}$. With this choice,

$$I \leq \exp\left(-\frac{|w|\sigma^2}{M^2}h\left(\frac{M\epsilon}{|w|\sigma^2}\right)\right).$$

This proves the desired inequality. ■

We can also derive concentration inequalities which are networked analogues of the Bernstein and Chernoff-Hoeffding inequalities and the proofs are in the appendix. These inequalities are used in the next section to provide learning guarantees.

Theorem 6 *Let $\{\xi_i\}_{i=1}^n$ be G-networked random variables with mean $\mathbb{E}[\xi_i] = \mu$, variance $\sigma^2(\xi_i) = \sigma^2$ (the variance condition is not needed for Inequality (8)), and satisfying $|\xi_i - \mu| \leq M$. Let w be a fractional matching of G and $|w| = \sum_{i=1}^n w_i$, then for all $\epsilon > 0$,*

$$\Pr\left(\frac{1}{|w|}\sum_{i=1}^{n} w_i\xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{|w|\epsilon}{2M}\log\left(1 + \frac{M\epsilon}{\sigma^2}\right)\right), \tag{6}$$

$$\Pr\left(\frac{1}{|w|}\sum_{i=1}^{n} w_i\xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{|w|\epsilon^2}{2(\sigma^2 + \frac{1}{3}M\epsilon)}\right), \text{ and} \tag{7}$$

$$\Pr\left(\frac{1}{|w|}\sum_{i=1}^{n} w_i\xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{|w|\epsilon^2}{2M^2}\right). \tag{8}$$

4.2. The FMN Scheme

According to Theorem 6, tighter bounds can be obtained by maximizing $|w|$ under the constraint that w is a fractional matching. Given a hypergraph, this can be achieved by solving the linear program (LP):

$$\max_{w} \quad \sum_{i=1}^{n} w_i \qquad \text{s.t.} \quad \forall i : w_i \geq 0 \quad \text{and} \quad \forall v \in V : \sum_{i: v \in e_i} w_i \leq 1.$$

A weight vector which makes the sum $\sum_{i=1}^{n} w_i$ maximum is called a *maximum fractional matching* of G. The optimal value of this linear program is called the *fractional matching number (FMN)* of the hypergraph G and is denoted by ν_G^*. That is, ν_G^* is defined as the sum of a maximum fractional matching vector.

There are very efficient LP solvers, including the simplex method which is efficient in practice, and the more recent interior-point methods (Boyd and Vandenberghe, 2004). The interior-point method solves an LP in $O(\mathsf{p}^2\mathsf{q})$ time, where p is the number of decision variables, and q is the number of constraints. In practice, usually every hyperedge does not connect many vertices and a vertex is not incident to many hyperedges, so these LPs are usually sparse. Almost all LP solvers perform significantly better for sparse LPs.

4.3. Improvement on EQW

Using Theorem 6, we can also improve Janson's inequalities for the EQW weighting scheme. Let w be a fractional matching and satisfy $w_1 = w_2 = \ldots = w_n$ (EQW). This requires that for all i, $0 < w_i \leq 1/\omega_G$ where $\omega_G = \max_{v \in V_G} |\{e : v \in e\}|$ is the maximum degree of G. Let $w_1 = w_2 = \ldots = w_n = 1/\omega_G$, we can get the following corollary.

Corollary 7 *Let $\{\xi_i\}_{i=1}^n$ be G-networked random variables with mean $\mathbb{E}[\xi_i] = \mu$, variance $\sigma^2(\xi_i) = \sigma^2$, and satisfying $|\xi_i - \mu| \leq M$. Then for all $\epsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\omega_G M} \log\left(1 + \frac{M\epsilon}{\sigma^2}\right)\right),$$

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\omega_G(\sigma^2 + \frac{1}{3}M\epsilon)}\right),$$

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} \xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\omega_G M^2}\right).$$

It is known that for every hypergraph, the maximum degree[3] is not larger than the fractional hyperedge-chromatic number, i.e., $\omega_G \leq \chi_G^*$. This fact generally ensures that the inequalities in Corollary 7 provide tighter bounds than those in Theorem 14. In addition, for any number $r \geq 1$, there exist hypergraphs G such that $\chi_G^*/\omega_G > r$, and hence the improvement of Corollary 7 over

3. Do not confuse the maximum degree ω_G with another concept - the maximum degree Δ of the corresponding dependency graphs (see (Janson, 2004)). It holds that $\chi_G^* \leq \Delta + 1$, but for any $c > 0$ there exist (infinitely many) hypergraphs such that $\chi_G^* > \omega_G + c$.

Theorem 14 can be arbitrarily large. As an example, we consider (truncated) projective planes[4] of order κ (see e.g., (Matousek and Nesetril, 1998)). It is known that there exists a projective plane of order κ whenever κ is a prime power. The fractional hyperedge-chromatic number of any subhypergraph of a projective plane is equal to its hyperedge number. For such datasets, Janson's inequalities (Theorem 14) fail to offer useful bounds while Corollary 7 provides significantly better bounds. Indeed, the maximum degree of the projective plane of order κ ($\kappa \geq 2$) is $\kappa + 1$, so $\chi^*_G/\omega_G = \kappa + 1/(\kappa+1) = \Omega(|E_G|^{1/2})$.

For different hypergraphs, it is possible that $n/\omega_G > \nu_G$, $n/\omega_G = \nu_G$ or $n/\omega_G < \nu_G$. Therefore, whether the IND scheme (i.e., the classical concentration bounds applied to independent subsets) provides tighter bounds than the EQW scheme depends on the hypergraph. However, both n/ω_G and ν_G are smaller than ν^*_G. Hence, the FMN scheme always gives better concentration bounds than the other two schemes.

4.4. U-statistics

Hoeffding (1948) presented concentration inequalities for U-statistics which is an important class of statistics on networked random variables. Above results also improve these concentration inequalities. As an example, we consider one-sample U-statistics.

Definition 8 (One-sample U-statistics) *Let $\{S_i\}_{i=1}^m$ be independent random variables. For $r \leq m$ consider a random variable of the form*

$$U = \frac{(m-r)!r!}{m!} \sum_{m,r} \xi(S_{i_1}, \ldots, S_{i_r})$$

where the sum $\sum_{m,r}$ is taken over all subset $\{i_1, \ldots, i_r\}$ of distinct positive integers not exceeding m. The random variable U is called a one-sample U-statistic.

If the function ξ is bounded, $|\xi - \mathbb{E}[\xi]| \leq M$, Hoeffding (1948) showed that for any $\epsilon > 0$,

$$\Pr\left(U - \mu \geq \epsilon\right) \leq \exp\left(-\frac{\lfloor \frac{m}{r} \rfloor \epsilon^2}{2M^2}\right)$$

where $\mu = \mathbb{E}[U]$. A corollary of our result shows that the operator $\lfloor\rfloor$ is not necessary

$$\Pr\left(U - \mu \geq \epsilon\right) \leq \exp\left(-\frac{m\epsilon^2}{2rM^2}\right).$$

To prove this inequality, we construct a hypergraph $G = (V_G, E_G)$ for a one-sample U-statistic. This hypergraph has m vertices, and $E_G = \{e \subseteq V_G \mid |e| = r\}$. We consider the independent random variables $\{S_i\}_{i=1}^m$ as the features of the vertices. The statistic U is an equally weighted sample mean of the networked random variables defined on the hyperedges. The inequality is proved by letting $n = \frac{m!}{(m-r)!r!}$ and $\omega_G = \frac{(m-1)!}{(m-r)!(r-1)!}$ in Corollary 7.

4. Projective planes and truncated projective planes are not only of theoretical interest. In fact, they are special cases of block designs as studied in the field of experimental design (Colbourn and Dimitz, 2010). This field studies what points in a feature space to measure to maximize certain experimental objectives such as diversity and independence of training data.

5. Application of new inequalities: Learning theory for networked examples

In this section, we use the results we obtained in the previous section to show generalization performance guarantees when learning from networked examples, making the same relaxed assumptions as in previous sections. We do this in the context of a specific framework (empirical risk minimization), but the same principles can be applied to many other paradigms of learning theory, e.g., structural risk minimization (Shawe-Taylor et al., 1998). We showed that the FMN weighting scheme provides clearly better properties than classical approaches.

The main goal of supervised learning is to learn a function $f : \mathbb{X} \mapsto \mathcal{Y}$ from a set of training examples $Z = \{z_i\}_{i=1}^n$ with $z_i = (x_i, y_i) \sim \rho$, and to predict labels for unseen examples.

For networked examples in hypergraph G, $x_i \in \mathbb{X}$ is the combination of the features of vertices in hyperedge e_i and $y_i \in \mathcal{Y}$ is the target value of hyperedge e_i. In this case, the dataset Z is called a G-networked sample.

We consider the least square regression which goal is to find a minimizer of the *expected risk*[5]

$$\mathcal{E}(f) = \int_{\mathcal{Z}} (f(x) - y)^2 \, \rho(x, y) dx dy.$$

That is,

$$f_{\rho,\mathcal{F}} = \arg\min_{f \in \mathcal{F}} \mathcal{E}(f),$$

where the minimization is taken over the set of all measurable functions $\mathcal{F}$. Unfortunately, since the probability distribution ρ is unknown, $f_{\rho,\mathcal{F}}$ cannot be computed directly. We find a good approximation of $f_{\rho,\mathcal{F}}$ from sample. The empirical risk minimization (ERM) principle (Vapnik, 1991) is to find the minimizer of empirical risk $\mathcal{E}_Z(f)$ in a properly selected hypothesis space $\mathcal{H}$, i.e.,

$$f_{Z,\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \left\{ \mathcal{E}_Z(f) = \frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \right\}.$$

The performance of the ERM approach is commonly measured by the *excess risk*

$$\mathcal{E}(f_{Z,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{F}}) = [\mathcal{E}(f_{Z,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{H}})] + [\mathcal{E}(f_{\rho,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{F}})],$$

where

$$f_{\rho,\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}(f).$$

We call the first part the *sample error* $\mathcal{E}_S(Z) := \mathcal{E}(f_{Z,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{H}})$, the second part the *approximation error* $\mathcal{E}_A(\mathcal{H}) := \mathcal{E}(f_{\rho,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{F}})$.

The approximation error is independent of the sample. In this paper, we concentrate on the sample error (we refer the readers who are interested in analyzing approximation errors to (Cucker and Zhou, 2007)). To this end, we need to choose a proper hypothesis space. The complexity of the hypothesis space is usually measured in terms of covering number (Zhou, 2002), entropy number (Tsuda, 1999), VC-dimension (Vapnik et al., 1994), etc. We use the covering numbers defined below to measure the capacity of our hypothesis space $\mathcal{H}$, and the hypothesis space $\mathcal{H}$ is usually chosen as a subset of $\mathcal{C}(\mathbb{X})$ which is a Banach space of continuous functions on a compact metric space $\mathbb{X}$ with

5. Similar results for general loss functions can be established, in this paper, we take the squared loss for the sake of simplicity.

the norm $\|f\|_\infty = \sup_{x\in\mathbb{X}} |f(x)|$. Our results can be extended to other complexity measures such as VC-dimensions, but it has been shown in Evgeniou and Pontil (1999) that VC-dimension is not suitable for real-valued function classes, while in this section the random variables are presented as real-valued functions, therefore, we measure the complexity of the hypothesis space by the uniform covering number.

Definition 9 (Covering number) *Let $\mathcal{H}$ be a metric space and $\tau > 0$. We define the* covering number *$N(\mathcal{H}, \tau)$ to be the minimal $\ell \in \mathbb{N}$ such that there exists ℓ disks in $\mathcal{H}$ with radius τ covering $\mathcal{H}$. When $\mathcal{H}$ is compact, this number is finite.*

Definition 10 (M-bounded functions) *Let $M > 0$ and ρ be a probability distribution on $\mathcal{Z}$. We say that a set $\mathcal{H}$ of functions from $\mathbb{X}$ to $\mathcal{Y}$ is* M-bounded *when*

$$\Pr_{(x,y)\sim\rho}\left(\sup_{f\in\mathcal{H}} |f(x) - y| \leq M\right) = 1.$$

Theorem 11 (Cucker and Zhou (2007)) *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(\mathbb{X})$, and Z be an i.i.d. sample. If $\mathcal{H}$ is M-bounded, then for all $\epsilon > 0$,*

$$\Pr\left(\mathcal{E}_S(Z) \geq \epsilon\right) \leq N\left(\mathcal{H}, \frac{\epsilon}{12M}\right) \exp\left(-\frac{n\epsilon^2}{300M^4}\right).$$

The above theorem provides learning guarantee for i.i.d. training samples.

Now, we consider three weighting schemes having different sample error bounds which are related to different important parameters of hypergraphs. The first two weighting schemes are straightforward, but from the upper bound point of view, they waste the information provided by the networked examples. The third weighting scheme reaches a better sample error bound via solving the linear program discussed in Section 4.1.

5.1. The EQW Scheme

Let us first consider the EQW weighting scheme that learns from a set of networked examples in the same way as if they were i.i.d., i.e., without weighting them as a function of the network structure. We can use Corollary 7 above to bound the sample error of EQW scheme:

Theorem 12 *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(\mathbb{X})$, and Z be a G-networked sample. If $\mathcal{H}$ is M-bounded, then for all $\epsilon > 0$,*

$$\Pr\left(\mathcal{E}_S(Z) \geq \epsilon\right) \leq N\left(\mathcal{H}, \frac{\epsilon}{12M}\right) \exp\left(-\frac{n\epsilon^2}{300\omega_G M^4}\right).$$

The result above shows that the bound of the sample error not only relies on the sample size but also depends on the maximum degree ω_G. That is, a larger sample may result in a poorer sample error bound since ω_G can also become larger. Remember that almost all previous works only deal with the EQW scheme, and their results depend on n/χ_G^*, so Theorem 12, which depends on n/ω_G, improves their results significantly.

5.2. The IND scheme

A straightforward idea to learn from a G-networked sample Z is to find a maximum subset $Z_I \subseteq Z$ of training examples that correspond to a matching in G. Due to our assumptions, such set is an i.i.d. sample. We can then perform algorithms on Z_I for learning, the function we obtain by the ERM principle is

$$f_{Z_I,\mathcal{H}} = \arg\min_{f\in\mathcal{H}} \left\{ \mathcal{E}_{Z_I}(f) = \frac{1}{|Z_I|} \sum_{z_i \in Z_I} (f(x_i) - y_i)^2 \right\}.$$

To bound the sample error of this weighting scheme, we can directly use Theorem 11, replacing n there by $|Z_I|$. As we discussed in Section 4, despite the IND scheme satisfies monotonicity, it is still far from optimal, due to the lack of efficiency. Besides, similar to the EQW scheme, the IND scheme may result in that networked datasets are not fully utilized.

5.3. The FMN scheme

We now consider the FMN weighting scheme proposed in Section 4.1. For a G-networked sample Z, we denote the FMN weighted sample $Z_{\nu^*} = \{(z_i, w_i)\}$ where $(w_i)_{i=1}^n$ is an FMN weight vector. Now we can define a new empirical risk on the FMN weighted sample Z_{ν^*} that

$$\mathcal{E}_{Z_{\nu^*}}(f) = \frac{1}{\nu^*} \sum_{i=1}^{n} w_i \left(f(x_i) - y_i\right)^2.$$

Later, we show that the empirical risk $\mathcal{E}_{Z_{\nu^*}}$ converges to the expected risk $\mathcal{E}(f)$ as ν^* tends to infinity for fixed f, then excess risk can be divided into two parts as follows

$$\mathcal{E}(f_{Z_{\nu^*},\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{F}}) = [\mathcal{E}(f_{Z_{\nu^*},\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{H}})] + [\mathcal{E}(f_{\rho,\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{F}})].$$

We use the probability inequalities with ν^* (see Theorem 6) to estimate the sample error $\mathcal{E}_S(Z_{\nu^*}) := \mathcal{E}(f_{Z_{\nu^*},\mathcal{H}}) - \mathcal{E}(f_{\rho,\mathcal{H}})$.

Theorem 13 *Let $\mathcal{H}$ be a compact and convex subset of $\mathcal{C}(\mathbb{X})$. If $\mathcal{H}$ is M-bounded, then for all $\epsilon > 0$,*

$$\Pr\left(\mathcal{E}_S(Z_{\nu^*}) \geq \epsilon\right) \leq N\left(\mathcal{H}, \frac{\epsilon}{12M}\right) \exp\left(-\frac{\nu^*\epsilon^2}{300M^4}\right).$$

6. Concluding remarks

In this paper, we considered the problem of learning from networked data. We proposed several schemes for weighting training examples that allow for using the available training data to a large extent while mitigating the dependency problem. In particular, the FMN weighting scheme allows for generalizing a large fraction of existing statistical learning theory. The weights in our weighting schemes can be computed efficiently. The presented theory forms a step towards a statistically sound theory for learning in networks.

We made here a weaker independence assumption. In order to analyze more real-world networked cases, one needs to build new models with more flexible assumptions than the classic i.i.d.

assumption and the weaker assumption we studied which is already more general but still is not sufficiently powerful to properly model a range of real-world scenarios. A first step in this direction would be to develop a measure to assess the strength of the dependency of the features between vertices and its influence on the learning task at hand.

To design active learning methods for structured data is another useful topic, i.e., to study query strategies to choose objects or examples in a network to perform experiments in order to learn a good predictor at minimal cost. It is also an interesting work to study the implications of the above theory and algorithms for other learning settings and tasks, e.g., cross-validation and bootstrapping (see e.g., (Liu et al., 1988)).

Acknowledgement

We thank Prof. Maurice Bruynooghe for detailed comments, Prof. Roger Wattenhofer for his support and Prof. Ding-Xuan Zhou for the fruitful discussions.

The work is supported partially by the National Natural Science Foundation of China (Grants No.11401524, 11531013, 11601375) and by ERC Starting Grant 240186 “MiGraNT: Mining Graphs and Networks: a Theory-based approach”.

References

Massih-Reza Amini and Nicolas Usunier. *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

George Bennett. Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association*, 57(297):33–45, 1962.

Sergei Bernstein. On a modification of Chebyshev's inequality and of the error formula of Laplace. *Ann. Sci. Inst. Sav. Ukraine, Sect. Math*, 1(4):38–49, 1924.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Richard C. Bradley. Basic properties of strong mixing conditions, a survey and some open questions. *Probability Surveys*, 2(2):107–144, 2005.

Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *The Annals of Mathematical Statistics*, 23(4):493–507, 1952.

Fan RK Chung and Linyuan Lu. *Complex Graphs and Networks*, volume 107. American mathematical society, 2006.

Charles J. Colbourn and Jeffrey H. Dimitz, editors. *Handbook of Combinatorial Designs*. CRC press, 2010.

Felipe Cucker and Ding-Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

Paul Erdos and László Lovász. Problems and results on 3-chromatic hypergraphs and some related questions. *Infinite and finite sets*, 10(2):609–627, 1975.

Theodoros Evgeniou and Massimiliano Pontil. On the v γ dimension for regression in reproducing kernel hilbert spaces. In *International Conference on Algorithmic Learning Theory*, pages 106–117. Springer, 1999.

Daan Fierens, Hendrik Blockeel, Maurice Bruynooghe, and Jan Ramon. Logical Bayesian networks and their relation to other probabilistic logical models. In *Inductive Logic Programming*, pages 121–135. Springer Berlin Heidelberg, 2005.

Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, volume 99, pages 1300–1309, 1999.

Michael R. Garey and David S. Johnson. *Computers and Intractibility, A Guide to the Theory of NP-Completeness*. W. H. Freeman Company, 1979.

Lise Getoor and Ben Taskar, editors. *Introduction to Statistical Relational Learning*. MIT press, 2007.

Zheng-Chu Guo and Lei Shi. Classification with non-iid sampling. *Mathematical and Computer Modelling*, 54(5):1347–1364, 2011.

Wassily Hoeffding. A class of statistics with asymptotically normal distributions. *Annals of Statistics*, 19(3):293–325, 1948.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Manfred Jaeger. Relational Bayesian networks. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 266–273. Morgan Kaufmann Publishers, 1997.

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

Johan Ludwig William Valdemar Jensen. Sur les fonctions convexes et les inégalités entre les valeurs moyennes. *Acta Mathematica*, 30(1):175–193, 1906.

David Knoke. *Economic networks*. John Wiley & Sons, 2014.

Samuel Leinhardt. *Social networks: A developing paradigm*. Elsevier, 2013.

Regina Y Liu et al. Bootstrap procedures under some non-iid models. *The Annals of Statistics*, 16 (4):1696–1708, 1988.

Tie-Yan Liu et al. Learning to rank for information retrieval. *Foundations and Trends® in Information Retrieval*, 3(3):225–331, 2009.

Jiri Matousek and Jaroslav Nesetril. *Invitation to Discrete Mathematics*. Oxford University Press, 1998.

Dharmendra S. Modha and Elias Masry. Minimum complexity regression estimation with weakly dependent observations. *Information Theory, IEEE Transactions on*, 42(6):2133–2145, 1996.

Michal Nykl, Karel Ježek, Dalibor Fiala, and Martin Dostal. Pagerank variants in the evaluation of citation networks. *Journal of Informetrics*, 8(3):683–692, 2014.

Liva Ralaivola and Massih-Reza Amini. Entropy-Based Concentration Inequalities for Dependent Variables. *Proceedings of the 32 nd International Conference on Machine Learning*, pages 1–9, 2015.

Liva Ralaivola, Marie Szafranski, and Guillaume Stempfel. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. *Journal of Machine Learning Research*, 11(July):1927–1956, 2010.

Ryan A Rossi, Luke K McDowell, David W Aha, and Jennifer Neville. Transforming graph data for statistical relational learning. *Journal of Artificial Intelligence Research*, 45(1):363–441, 2012.

Edward R Scheinerman and Daniel H Ullman. *Fractional graph theory: a rational approach to the theory of graphs*. Courier Corporation, 2011.

John Shawe-Taylor, Peter L. Bartlett, Robert C. Williamson, and Martin Anthony. Structural risk minimization over data dependent hierarchies. *Information Theory, IEEE Transactions on*, 44(5): 1926–1940, September 1998.

Hongwei Sun and Qiang Wu. Regularized least square regression with dependent samples. *Advances in Computational Mathematics*, 32(2):175–189, 2010.

Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P Tsafou, et al. String v10: protein–protein interaction networks, integrated over the tree of life. *Nucleic acids research*, 43(D1):D447–D452, 2014.

Koji Tsuda. Optimal hyperplane classifier based on entropy number bound. In *Artificial Neural Networks, 1999. ICANN 99. Ninth International Conference on (Conf. Publ. No. 470)*, volume 1, pages 419–424, 1999.

Feige Uriel, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy. Approximating clique is almost NP-complete. In *Foundations of Computer Science, 1991. Proceedings., 32nd Annual Symposium on. IEEE*, pages 2–12, 1991.

Nicolas Usunier, Massih-reza Amini, and Patrick Gallinari. Generalization error bounds for classifiers trained with interdependent data. In *Advances in Neural Information Processing Systems 18 (NIPS 2005)*, pages 1369–1376. MIT Press, 2006.

Vladimir Vapnik. Principles of risk minimization for learning theory. *Advances in Neural Information Processing Systems 4 (NIPS 1991)*, pages 831–838, 1991.

Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

Tao Wang, Jennifer Neville, Brian Gallagher, and Tina Eliassi-Rad. Correcting bias in statistical tests for network classifier evaluation. In *Proceedings of ECML/PKDD*, volume 6913 of *LNCS*, pages 506–521, 2011.

Ding-Xuan Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Appendix A. EQW and Janson's Inequalities

Janson (2004) showed inequalities on the gap between the expected value μ and the unweighted average of n networked random variables ξ_i.

Theorem 14 (Janson (2004)) *Let $\{\xi_i\}_{i=1}^n$ be G-networked random variables with mean $\mathbb{E}[\xi_i] = \mu$, variance $var(\xi_i) = \sigma^2$ and satisfying $|\xi_i - \mu| \leq M$. Then for all $\epsilon > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^n \xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{n\epsilon^2}{2\chi_G^* M^2}\right),$$
$$\Pr\left(\frac{1}{n}\sum_{i=1}^n \xi_i - \mu \geq \epsilon\right) \leq \exp\left(-\frac{8n\epsilon^2}{25\chi_G^*(\sigma^2 + M\epsilon/3)}\right).$$

Remark Note that the second inequality above is a Bernstein-type inequality, but it cannot degenerate to the original Bernstein inequality when the random variables are independent. In Section 4.3, we show an improvement on these inequalities, which not only removes the unnecessary coefficient in the second inequality and then makes the inequality can degenerate to the original independent case when random variables are independent, but also replaces χ_G^* by another parameter which is much smaller than χ_G^* in many cases.

Appendix B. Proofs Omitted from Section 4

In this part, we prove Theorem 6.

Lemma 15 *Let $\beta = (\beta_i)_{i=1}^k \in \mathbb{R}_+^k$ such that $\sum_{i=1}^k \beta_i \leq 1$. Then, the function $g(t)$ with $t = (t_i)_{i=1}^k \in \mathbb{R}_+^k$ defined by $g(t) = \prod_{i=1}^k t_i^{\beta_i}$, is concave.*

Proof We prove by showing that its Hessian matrix $\nabla^2 g(t)$ is negative semidefinite. $\nabla^2 g(t)$ is given by

$$\frac{\partial^2 g(t)}{\partial t_i^2} = \frac{\beta_i(\beta_i - 1)g(t)}{t_i^2}, \qquad \frac{\partial^2 g(t)}{\partial t_i \partial t_j} = \frac{\beta_i \beta_j g(t)}{t_i t_j},$$

and can be expressed as

$$\nabla^2 g(t) = \left(qq^{\mathrm{T}} - \mathbf{diag}(\beta_1/t_1^2, \ldots, \beta_n/t_n^2)\right) g(t)$$

where $q = [q_1, \ldots, q_k]$ and $q_i = \beta_i/t_i$. We must show that $\nabla^2 g(t) \preceq 0$, i.e., that

$$u^{\mathrm{T}} \nabla^2 g(t) u = \left(\left(\sum_{i=1}^k \beta_i u_i/t_i\right)^2 - \sum_{i=1}^k \beta_i u_i^2/t_i^2\right) g(t) \leq 0$$

for all $u \in \mathbb{R}^k$. Because $g(t) \geq 0$ for all t, we only need to prove

$$\left(\sum_{i=1}^k \beta_i u_i/t_i\right)^2 - \sum_{i=1}^k \beta_i u_i^2/t_i^2 \leq 0.$$

Since β_i is positive for every i and $\sum_{i=1}^k \beta_i \leq 1$, we define a random variable ξ with probability $P(\xi = u_i/t_i) = \beta_i$ and $P(\xi = 0) = 1 - \sum_{i=1}^k \beta_i$. From basic probability theory, we have

$$\left(\sum_{i=1}^k \beta_i u_i/t_i\right)^2 = (\mathbb{E}[\xi])^2 \leq \mathbb{E}\left[\xi^2\right] = \sum_{i=1}^k \beta_i u_i^2/t_i^2.$$

■

Lemma 16 *Let $(\xi_i)_{i=1}^n$ be G-networked random variables with mean $\mathbb{E}[\xi_i] = \mu$ and variance $\sigma^2(\xi_i) = \sigma^2$, such that $|\xi_i - \mu| \leq M$. Let $w = (w_i)_{i=1}^n$ be a vertex-bounded weight vector for G and let $|w| = \sum_i w_i$, then for all $\epsilon > 0$,*

$$\Pr\left(\sum_{i=1}^n w_i\left(\xi_i - \mu\right) \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon}{2M}\log(1 + \frac{M\epsilon}{|w|\sigma^2})\right),$$
$$\Pr\left(\sum_{i=1}^n w_i\left(\xi_i - \mu\right) \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2(|w|\sigma^2 + \frac{1}{3}M\epsilon)}\right),$$
$$\Pr\left(\sum_{i=1}^n w_i\left(\xi_i - \mu\right) \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2}{2|w|M^2}\right).$$

Proof Without loss of generality, we assume $\mu = 0$. The first inequality follows from Lemma 5 and the inequality

$$h(a) \geq \frac{a}{2}\log(1 + a),\ \forall a \geq 0.$$

The second inequality follows from Lemma 5 and the inequality

$$h(a) \geq \frac{3a^2}{6 + 2a},\ \forall a \geq 0.$$

To prove the third inequality, we use Lemma 4. As the exponential function is convex and $-M \leq \xi_i \leq M$, there holds

$$e^{c\xi_i} \leq \frac{c\xi_i - (-cM)}{2cM}e^{cM} + \frac{cM - c\xi_i}{2cM}e^{-cM}.$$

It follows from the assumption $\mu = 0$ and the Taylor expansion for the exponential function that

$$\begin{aligned}\mathbb{E}\left[e^{c\xi_i}\right] \leq & \frac{1}{2}e^{-cM} + \frac{1}{2}e^{cM} = \frac{1}{2}\sum_{p=0}^{+\infty}\frac{(-cM)^p}{p!} + \frac{1}{2}\sum_{p=0}^{+\infty}\frac{(cM)^p}{p!} = \sum_{p=0}^{+\infty}\frac{(cM)^{2p}}{(2p)!} \\ = & \sum_{p=0}^{+\infty}\frac{((cM)^2/2)^p}{p!}\prod_{j=1}^{p}\frac{1}{2j-1} \leq \sum_{p=0}^{+\infty}\frac{((cM)^2/2)^p}{p!} = \exp((cM)^2/2).\end{aligned}$$

This, together with Lemma 4, implies

$$\begin{aligned}\Pr\left(\sum_{i=1}^{n} w_i\left(\xi_i-\mu\right)\geq\epsilon\right) &= \Pr\left(\exp\left(c\sum_{i=1}^{n} w_i\xi_i\right)\geq e^{c\epsilon}\right)\\ &\leq \exp\left(-c\epsilon+\mathbb{E}\left[c\sum_{i=1}^{n} w_i\xi_i\right]\right)\\ &\leq \exp\left(-c\epsilon+|w|(cM)^2/2\right).\end{aligned}$$

Choose $c=\epsilon/(|w|M^2)$. Then, $\Pr\left(\sum_{i=1}^{n} w_i\left(\xi_i-\mu\right)\geq\epsilon\right)\leq\exp\left(-\frac{\epsilon^2}{2|w|M^2}\right)$. ■

Now we are ready to prove Theorem 6.

Proof [proof of Theorem 6] We apply Lemma 16 to the variables $\xi_i'=\xi_i/|w|$ which satisfy $|\xi_i'-\mathbb{E}[\xi_i']|\leq M/|w|, \sigma^2(\xi_i')=\sigma^2/|w|^2$. ■

Appendix C. Proofs Omitted from Section 5

In this part we prove Theorem 13. We first give some lemmas which are extended versions of lemmas that were used before to establish the sample error bounds for i.i.d. samples. In particular, some ideas were borrowed from (Cucker and Zhou, 2007). For any function $f\in\mathcal{H}$, we define the defect function $\mathcal{D}_{Z_{\nu^*}}(f)=\mathcal{E}(f)-\mathcal{E}_{Z_{\nu^*}}(f)$, the difference between the expected risk of f and the empirical risk of f on the FMN weighted sample Z_{ν^*}.

Lemma 17 *Let $M>0$ and let $f:\mathbb{X}\mapsto\mathcal{Y}$ be M-bounded. Then for all $\epsilon>0$,*

$$\Pr\left(\mathcal{D}_{Z_{\nu^*}}(f)\geq-\epsilon\right)\geq 1-\exp\left(\frac{\nu^*\epsilon^2}{2M^4}\right).$$

Proof Note that $\Pr\left(\mathcal{D}_{Z_{\nu^*}}(f)\geq-\epsilon\right)=\Pr\left(\mathcal{E}_{Z_{\nu^*}}(f)-\mathcal{E}(f)\leq\epsilon\right)$. This lemma then follows directly from Inequality (8) in Theorem 6 by taking $\xi_i=(f(x_i)-y_i)^2$ satisfying $|\xi_i|\leq M^2$ when f is M-bounded. ■

To present Lemma 19 and 20, we first define full measure sets.

Definition 18 (full measure set) *A set $U\subseteq\mathcal{Z}$ is full measure for distribution ρ over $\mathcal{Z}$ if $\Pr_{z\sim\rho}(z\in U)=1$.*

Lemma 19 *If for $j=1,2$, $|f_j(x)-y|\leq M$ on a full measure set $U\subseteq\mathcal{Z}$ then, for all $Z\in U^n$*

$$|\mathcal{D}_{Z_{\nu^*}}(f_1)-\mathcal{D}_{Z_{\nu^*}}(f_2)|\leq 4M\|f_1-f_2\|_\infty.$$

Proof Because

$$(f_1(x)-y)^2-(f_2(x)-y)^2=(f_1(x)+f_2(x)-2y)(f_1(x)-f_2(x)),$$

we have

$$\begin{aligned}|\mathcal{E}(f_1)-\mathcal{E}(f_2)| &= \left|\int_{\mathcal{Z}}\rho(z)(f_1(x)+f_2(x)-2y)(f_1(x)-f_2(x))\mathrm{d}z\right| \\ &\leq \int_{\mathcal{Z}}\rho(z)|(f_1(x)-y)+(f_2(x)-y)|\|f_1-f_2\|_\infty\mathrm{d}z \\ &\leq 2M\|f_1-f_2\|_\infty.\end{aligned}$$

For $Z\in U^n$, we have

$$\begin{aligned}|\mathcal{E}_{Z_{\nu^*}}(f_1)-\mathcal{E}_{Z_{\nu^*}}(f_2)| &= \frac{1}{\nu^*}\sum_{i=1}^n w_i(f_1(x_i)+f_2(x_i)-2y_i)(f_1(x_i)-f_2(x_i) \\ &\leq \frac{1}{\nu^*}\sum_{i=1}^n w_i|(f_1(x_i)-y_i)+(f_2(x_i)-y_i)|\|(f_1-f_2\|_\infty \\ &\leq 2M\|f_1-f_2\|_\infty.\end{aligned}$$

Thus,

$$|\mathcal{D}_{Z_{\nu^*}}(f_1)-\mathcal{D}_{Z_{\nu^*}}(f_2)|=|\mathcal{E}(f_1)-\mathcal{E}_{Z_{\nu^*}}(f_1)-\mathcal{E}(f_2)+\mathcal{E}_{Z_{\nu^*}}(f_2)|\leq 4M\|f_1-f_2\|_\infty.$$

■

Lemma 20 *Let $\mathcal{H}$ be a compact M-bounded subset of $\mathcal{C}(\mathbb{X})$. Then, for all $\epsilon>0$,*

$$\Pr\left(\sup_{f\in\mathcal{H}}\mathcal{D}_{Z_{\nu^*}}(f)\leq\epsilon\right)\geq 1-N\left(\mathcal{H},\frac{\epsilon}{8M}\right)\exp\left(-\frac{\nu^*\epsilon^2}{8M^4}\right).$$

Proof Let $\{f_j\}_{j=1}^\ell\subset\mathcal{H}$ with $\ell=N\left(\mathcal{H},\frac{\epsilon}{4M}\right)$ such that $\mathcal{H}$ is covered by disks D_j centered at f_j with radius $\frac{\epsilon}{4M}$. Let U be a full measure set on which $\sup_{f\in\mathcal{H}}|f(x)-y|\leq M$. Then for all $Z\in U^n$ and for all $f\in D_j$, according to Lemma 19, we have

$$|\mathcal{D}_{Z_{\nu^*}}(f)-\mathcal{D}_{Z_{\nu^*}}(f_j)|\leq 4M\|f-f_j\|_\infty\leq 4M\frac{\epsilon}{4M}=\epsilon.$$

Consequently,

$$\sup_{f\in D_j}\mathcal{D}_{Z_{\nu^*}}(f)\geq 2\epsilon\Rightarrow\mathcal{D}_{Z_{\nu^*}}(f_j)\geq\epsilon.$$

Then we conclude that, for $j=1,\cdots,\ell$,

$$\Pr\left(\sup_{f\in D_j}\mathcal{D}_{Z_{\nu^*}}(f)\geq 2\epsilon\right)\leq\Pr\left(\mathcal{D}_{Z_{\nu^*}}(f_j)\geq\epsilon\right)\leq\exp\left(-\frac{\nu^*\epsilon^2}{2M^4}\right).$$

The last inequality follows from Inequality (8) in Theorem 6 by taking $\xi_i=-(f_j(x_i)-y_i)^2$. In addition, one can easily see that

$$\sup_{f\in\mathcal{H}}\mathcal{D}_{Z_{\nu^*}}(f)\geq\epsilon\Leftrightarrow\exists j\leq\ell:\sup_{f\in D_j}\mathcal{D}_{Z_{\nu^*}}(f)\geq\epsilon$$

and from the fact that the probability of a union of events is bounded by the sum of the probabilities of these events it follows that

$$\Pr\left(\sup_{f\in\mathcal{H}} \mathcal{D}_{Z_{\nu^*}}(f) \geq \epsilon\right) \leq \sum_{j=1}^{\ell} \Pr\left(\sup_{f\in D_j} \mathcal{D}_{Z_{\nu^*}}(f) \geq \epsilon\right) \leq \ell \exp\left(-\frac{\nu^*\epsilon^2}{8M^4}\right).$$

This completes the proof. ■

Lemma 21 *Suppose networked random variables $(\xi_i)_{i=1}^n$ satisfy that for all i, $\mathbb{E}[\xi_i] = \mu \geq 0$, and $|\xi_i - \mu| \leq B$ almost everywhere. Let $(w_i)_{i=1}^n$ be any FMN weight vector. If $\mathbb{E}\left[\xi_i^2\right] \leq c\mu$, then for every $\epsilon > 0$ and $0 < \alpha \leq 1$, there holds*

$$\Pr\left(\frac{\mu - \frac{1}{\nu^*}\sum_{i=1}^n w_i\xi_i}{\sqrt{\mu+\epsilon}} > \alpha\sqrt{\epsilon}\right) \leq \exp\left(-\frac{\alpha^2\nu^*\epsilon}{2c + \frac{2}{3}B}\right).$$

Proof We apply Inequality (7) in Theorem 6 by substituting the ξ_i in Inequality (7) with $\xi_i/\sqrt{\mu+\epsilon}$, the ϵ in Inequality (7) with $\alpha\sqrt{\epsilon}$, the M in Inequality (7) with $B/\sqrt{\mu+\epsilon}$ and the $|w|$ in Inequality (7) with ν^*. We get

$$\Pr\left(\frac{\mu - \frac{1}{\nu^*}\sum_{i=1}^n w_i\xi_i}{\sqrt{\mu+\epsilon}} > \alpha\sqrt{\epsilon}\right) \leq \exp\left(-\frac{\alpha^2\nu^*\epsilon}{2(\sigma^2 + B\alpha\sqrt{\epsilon}/3\sqrt{\mu+\epsilon})}\right),$$

where $\sigma^2 = \mathbb{E}\left[(\xi_i/\sqrt{\mu+\epsilon})^2\right] \leq c\mu/(\mu+\epsilon)$. The lemma then follows from observing that $c\mu/(\mu+\epsilon) \leq c$ (as $\mu \geq 0$ and $\epsilon > 0$) and $B\alpha\sqrt{\epsilon}/3\sqrt{\mu+\epsilon} \leq B/3$ (as $\mu \geq 0$, $\epsilon \geq 0$ and $0 < \alpha \leq 1$). ■

Lemma 21 can also be extended to families of functions as follows.

Lemma 22 *Let $\mathcal{G}$ be a set of functions on $\mathcal{Z}$ and $c > 0$ such that, for each $g \in \mathcal{G}$, $\mathbb{E}[g] \geq 0$, $\mathbb{E}\left[g^2\right] \leq c\mathbb{E}[g]$ and $|g - \mathbb{E}[g]| \leq B$ almost everywhere. Let $(w_i)_{i=1}^n$ be any FMN weight vector. Then for every $\epsilon > 0$ and $0 < \alpha \leq 1$, we have*

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\mathbb{E}[g] - \frac{1}{\nu^*}\sum_{i=1}^n w_i g(z_i)}{\sqrt{\mathbb{E}[g]+\epsilon}} \geq 4\alpha\sqrt{\epsilon}\right) \leq N(\mathcal{G}, \alpha\epsilon)\exp\left(-\frac{\alpha^2\nu^*\epsilon}{2c + \frac{2}{3}B}\right).$$

Proof Let $\{g_j\}_{j=1}^J \subset \mathcal{G}$ with $J = N(\mathcal{G}, \alpha\epsilon)$ be such that $\mathcal{G}$ is covered by balls in $\mathcal{C}(\mathcal{Z})$ centered at g_j with radius $\alpha\epsilon$.

Applying Lemma 21 to $\xi_i = g_j(z_i)$ for each j, we have

$$\Pr\left(\frac{\mathbb{E}[g_j] - \frac{1}{\nu^*}\sum_{i=1}^n w_i g_j(z_i)}{\sqrt{\mathbb{E}[g_j]+\epsilon}} \geq \alpha\sqrt{\epsilon}\right) \leq \exp\left(-\frac{\alpha^2\nu^*\epsilon}{2c + \frac{2}{3}B}\right).$$

For each $g \in \mathcal{G}$, there is some j such that $||g - g_j||_{\mathcal{C}(\mathcal{Z})} \leq \alpha\epsilon$. Then $|\frac{1}{\nu^*}\sum_{i=1}^n g(z_i) - \frac{1}{\nu^*}\sum_{i=1}^n w_i g_j(z_i)|$ and $|\mathbb{E}[g] - \mathbb{E}[g_j]|$ are both bounded by $\alpha\epsilon$. Hence, as $\frac{\sqrt{\epsilon}}{\sqrt{\epsilon+\mathbb{E}[g]}} \leq 1$,

$$\frac{|\frac{1}{\nu^*}\sum_{i=1}^n g(z_i) - \frac{1}{\nu^*}\sum_{i=1}^n g_j(z_i)|}{\sqrt{\mathbb{E}[g]+\epsilon}} \leq \alpha\sqrt{\epsilon}$$

and

$$\frac{|\mathbb{E}[g]-\mathbb{E}[g_j]|}{\sqrt{\mathbb{E}[g]+\epsilon}} \leq \alpha\sqrt{\epsilon}.$$

The latter implies that

$$\begin{aligned}\mathbb{E}[g_j]+\epsilon &= \mathbb{E}[g_j]-\mathbb{E}[g]+\mathbb{E}[g]+\epsilon \leq \alpha\sqrt{\epsilon}\sqrt{\mathbb{E}[g]+\epsilon}+(\mathbb{E}[g]+\epsilon)\\ &\leq \sqrt{\epsilon}\sqrt{\mathbb{E}[g]+\epsilon}+(\mathbb{E}[g]+\epsilon) \leq 2(\mathbb{E}[g]+\epsilon).\end{aligned}$$

It follows that $\sqrt{\mathbb{E}[g_j]+\epsilon} \leq 2\sqrt{\mathbb{E}[g]+\epsilon}$. We have thus seen that $\frac{\mathbb{E}[g]-\frac{1}{\nu^*}\sum_{i=1}^n g(z_i)}{\sqrt{\mathbb{E}[g]+\epsilon}} \geq 4\alpha\sqrt{\epsilon}$ implies $\frac{\mathbb{E}[g_j]-\frac{1}{\nu^*}\sum_{i=1}^n w_i g_j(z_i)}{\sqrt{\mathbb{E}[g]+\epsilon}} \geq 2\alpha\sqrt{\epsilon}$ and hence $\frac{\mathbb{E}[g_j]-\frac{1}{\nu^*}\sum_{i=1}^n w_i g_j(z_i)}{\sqrt{\mathbb{E}[g_j]+\epsilon}} \geq \alpha\sqrt{\epsilon}$. Therefore,

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\mathbb{E}[g]-\frac{1}{\nu^*}\sum_{i=1}^n w_i g(z_i)}{\sqrt{\mathbb{E}[g]+\epsilon}} \geq 4\alpha\sqrt{\epsilon}\right) \leq \sum_{j=1}^J \Pr\left(\frac{\mathbb{E}[g_j]-\frac{1}{\nu^*}\sum_{i=1}^n w_i g_j(z_i)}{\sqrt{\mathbb{E}[g_j]+\epsilon}} \geq \alpha\sqrt{\epsilon}\right)$$

which is bouned by $J\cdot\exp\left(-\frac{\alpha^2\nu^*\epsilon}{2c+\frac{2}{3}B}\right)$. ■

Let $\mathcal{L}^2_\rho(\mathbb{X})$ be a Banach space with the norm $\|f\|_{\mathcal{L}^2_\rho(\mathbb{X})} = \left(\int_{\mathbb{X}} |f(x)^2|\rho_{\mathbb{X}}(x)\mathrm{d}x\right)^{\frac{1}{2}}$. where $\rho_{\mathbb{X}}(x) = \prod_{i=1}^k x^{(i)}$. We define the error in $\mathcal{H}$ of a function $f\in\mathcal{H}$,

$$\mathcal{E}_{\mathcal{H}}(f) = \mathcal{E}(f) - \mathcal{E}(f_{\rho,\mathcal{H}})$$

which is always nonnegative.

Lemma 23 *Let $\mathcal{H}$ be a convex subset of $\mathcal{C}(\mathbb{X})$ such that $f_{\rho,\mathcal{H}}$ exists. Then $f_{\rho,\mathcal{H}}$ is unique as an element in $\mathcal{L}^2_\rho(\mathbb{X})$ and for all $f\in\mathcal{H}$,*

$$\int_{\mathbb{X}} (f_{\rho,\mathcal{H}}(x)-f(x))^2\rho_{\mathbb{X}}(x)dx \leq \mathcal{E}_{\mathcal{H}}(f).$$

In particular, if $\rho_{\mathbb{X}}(x)$ is not degenerate then $f_{\rho,\mathcal{H}}$ is unique in $\mathcal{H}$.

Proof The proof can be found in (Cucker and Zhou, 2007) (Lemma 3.16). ■

Proof of Theorem 13 For every function $f\in\mathcal{H}$, we define a function

$$g_f(x,y) = (f(x)-y)^2 - (f_{\rho,\mathcal{H}}(x)-y)^2.$$

We define $\mathcal{G}$ as the set of all functions g_f with $f\in\mathcal{H}$. For any function $g_f\in\mathcal{G}$, we have

$$\mathbb{E}_{z\sim\rho}[g_f] = \mathcal{E}_{\mathcal{H}}(f) \geq 0. \tag{9}$$

We first show that the two preconditions of Lemma 22 are true (for $B=2M^2$ and $c=4M^2$):

1. $|g_f - \mathbb{E}_{z\sim\rho}[g_f]| \leq 2M^2$

2. $\mathbb{E}_{z\sim\rho}\left[g_f^2\right] \leq 4M^2\mathbb{E}_{z\sim\rho}[g_f]$.

First, since $\mathcal{H}$ is M-bounded, we have that $-M^2 \leq g_f(z) \leq M^2$ holds almost everywhere. It follows that $|g_f - \mathbb{E}_{z\sim\rho}[g_f]| \leq 2M^2$ holds almost everywhere. This is the first precondition above. Second, one can easily see that

$$g_f(z) = (f(x) - f_{\rho,\mathcal{H}}(x))[(f(x) - y) + (f_{\rho,\mathcal{H}}(x) - y)].$$

It follows that $|g_f(z)| \leq 2M|f(x) - f_{\rho,\mathcal{H}}(x)|$ holds almost everywhere. Then, $\mathbb{E}_{z\sim\rho}\left[g_f^2\right] \leq 4M^2\mathbb{E}_{x\sim\rho_{\mathbb{X}}}\left[(f(x) - f_{\rho,\mathcal{H}}(x))^2\right] = 4M^2\int_{\mathbb{X}}(f(x) - f_{\rho,\mathcal{H}}(x))^2\rho_{\mathbb{X}}(x)\mathrm{d}x$. Together with Lemma 23 this implies that $\mathbb{E}_{z\sim\rho}\left[g_f^2\right] \leq 4M^2\mathcal{E}_{\mathcal{H}}(f) = c\mathbb{E}_{z\sim\rho}[g_f]$ with $c = 4M^2$. Hence, all the conditions of Lemma 22 hold and we get that for every $\epsilon > 0$ and $0 < \alpha \leq 1$,

$$\Pr\left(\sup_{g\in\mathcal{G}} \frac{\mathbb{E}[g] - \frac{1}{\nu^*}\sum_{i=1}^n w_i g(z_i)}{\sqrt{\mathbb{E}[g_j] + \epsilon}} \geq 4\alpha\sqrt{\epsilon}\right) \leq N(\mathcal{G}, \alpha\epsilon)\exp\left(-\frac{\alpha^2\nu^*\epsilon}{2.4M^2 + \frac{2}{3}2M^2}\right). \tag{10}$$

Remind from Equation (9) that $\mathbb{E}[g_f] = \mathcal{E}_{\mathcal{H}}(f)$. We also define

$$\mathcal{E}_{\mathcal{H},Z_{\nu^*}}(f) = \frac{1}{\nu^*}\sum_{i=1}^n w_i g_f(z_i) = \frac{1}{\nu^*}\sum_{i=1}^n w_i(f(x) - y)^2 - \frac{1}{\nu^*}\sum_{i=1}^n w_i(f_{\rho,\mathcal{H}}(x) - y)^2$$

Furthermore, we take $\alpha = \sqrt{2}/8$. Substituting all these into Inequality (10) we get

$$\forall\epsilon > 0, \Pr\left(\sup_{f\in\mathcal{H}} \frac{\mathcal{E}_{\mathcal{H}}(f) - \mathcal{E}_{\mathcal{H},Z_{\nu^*}}(f)}{\sqrt{\mathcal{E}_S(f) + \epsilon}} \geq 4\frac{\sqrt{2}}{8}\sqrt{\epsilon}\right) \leq N\left(\mathcal{G}, \frac{\sqrt{2}}{8}\epsilon\right)\exp\left(-\frac{\left(\frac{\sqrt{2}}{8}\right)^2\nu^*\epsilon}{28M^2/3}\right).$$

As this holds for the supremum over f, it also holds for $f = f_{Z_{\nu^*},\mathcal{H}}$:

$$\forall\epsilon > 0, \Pr\left(\frac{\mathcal{E}_{\mathcal{H}}(f_{Z_{\nu^*},\mathcal{H}}) - \mathcal{E}_{\mathcal{H},Z_{\nu^*}}(f_{Z_{\nu^*},\mathcal{H}})}{\sqrt{\mathcal{E}_{\mathcal{H}}(f_{Z_{\nu^*},\mathcal{H}}) + \epsilon}} \geq \sqrt{\frac{\epsilon}{2}}\right) \leq N\left(\mathcal{G}, \frac{\sqrt{2}}{8}\epsilon\right)\exp\left(-\frac{\nu^*\epsilon}{896M^2/3}\right).$$

The definition of $f_{Z_{\nu^*},\mathcal{H}}$ tells us that $\mathcal{E}_S(Z_{\nu^*}) = \mathcal{E}_{\mathcal{H}}(f_{Z_{\nu^*},\mathcal{H}})$ and $\mathcal{E}_{\mathcal{H},Z_{\nu^*}}(f_{Z_{\nu^*},\mathcal{H}}) \leq 0$. It follows that (we also upper-bound $896/3$ by 300)

$$\forall\epsilon > 0, \Pr\left(\frac{\mathcal{E}_S(Z_{\nu^*})}{\sqrt{\mathcal{E}_S(Z_{\nu^*}) + \epsilon}} \geq \sqrt{\frac{\epsilon}{2}}\right) \leq N\left(\mathcal{G}, \frac{\sqrt{2}}{8}\epsilon\right)\exp\left(-\frac{\nu^*\epsilon}{300M^2}\right).$$

It is easy to see that $\mathcal{E}_S(Z_{\nu^*}) \geq \epsilon$ implies $\frac{\mathcal{E}_S(Z_{\nu^*})}{\sqrt{\mathcal{E}_S(Z_{\nu^*})+\epsilon}} \geq \sqrt{\frac{\epsilon}{2}}$, so

$$\forall\epsilon > 0, \Pr\left(\mathcal{E}_S(Z_{\nu^*}) \geq \epsilon\right) \leq N\left(\mathcal{G}, \frac{\sqrt{2}}{8}\epsilon\right)\exp\left(-\frac{\nu^*\epsilon}{300M^2}\right).$$

Finally, the inequality $\|g_{f_1} - g_{f_2}\|_{\mathcal{C}(\mathcal{Z})} = \|f_1(x) - f_2(x)[(f_1(x) - y) + (f_2(x) - y)]\|_{\mathcal{C}(\mathcal{Z})} \leq 2M\|f_1 - f_2\|_{\mathcal{C}(\mathbb{X})}$, tells us that

$$N(\mathcal{G}, \frac{\sqrt{2}\epsilon}{8}) \leq N(\mathcal{H}, \frac{\sqrt{2}\epsilon}{16M}) \leq N(\mathcal{H}, \frac{\epsilon}{12M}).$$

This completes our proof. ■