Entropy-Based Concentration Inequalities for Dependent Variables

Liva Ralaivola LIVA.RALAIVOLA@LIF.UNIV-MRS.FR
QARMA, LIF, CNRS, Aix-Marseille University, F–13288 Marseille cedex 9, France

Massih-Reza Amini MASSIH-REZA.AMINI@IMAG.FR
AMA, LIG, CNRS, University Grenoble Alpes, Centre Equation 4, BP 53, F–38041 Grenoble Cedex 9, France

Abstract

We provide new concentration inequalities for functions of dependent variables. The work extends that of Janson (2004), which proposes concentration inequalities using a combination of the Laplace transform and the idea of fractional graph coloring, as well as many works that derive concentration inequalities using the entropy method (see, e.g., (Boucheron et al., 2003)). We give inequalities for *fractionally sub-additive* and *fractionally self-bounding functions*. In the way, we prove a new Talagrand concentration inequality for fractionally sub-additive functions of dependent variables. The results allow us to envision the derivation of generalization bounds for various applications where dependent variables naturally appear, such as in bipartite ranking.

1. Introduction

We present new concentration inequalities for specific functions of possibly dependent random variables. The approach that we advocate is based on the entropy method and the idea of breaking up the dependencies between random variables thanks to a graph coloring approach. Having these results at hand allows us to envision the study of the generalization properties of predictors trained over interdependent data for which a suitable dependency structure exist. As discussed by Amini & Usunier (2015), this structure could be naturally related to the *dependency graph* of the data, or it could be obtained a posteriori from a transformation that reduces a general learning problem to a more simple case, e.g. some reductions of multiclass classification problems to binary classification problems.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37.

Related Works. Learning with interdependent data is a topic that has received quite interest over the past few years; from a theoretical point of view, it ultimately pertains to the availability of concentration inequalities designed to account for the dependencies at hand. Among the prominent works that address this problem are a series of contributions on learning from mixing processes, where the dependencies within a sequence of random variables decreases over time (Yu, 1994; Karandikar & Vidyasagar, 2002; Kontorovich & Ramanan, 2008; Mohri & Rostamizadeh, 2008; 2009; Samson, 2000; Steinwart & Christmann, 2010). Another line of research within this field, is based on the idea of graph coloring, designed to divide a graph into sets of *independent* sets, and considers subsets of independent random variables deduced from the graph, linking these variables. By mixing the idea of graph coloring with the Laplace transform, Hoeffding-like concentration inequalities for the sum of dependent random variables were proposed by Janson (2004). Usunier et al. (2006) later extended this result to provide a generalization of the bounded differences inequality of McDiarmid (1989) to the case of interdependent random variables. This extension then paved the way for the definition of the *fractional Rademacher complexity* that generalizes the idea of Rademancher complexity and allows one to derive generalization bounds for scenarios where the training data are made of dependent data. The *Chromatic PAC-Bayes* bound proposed by Ralaivola et al. (2009; 2010) is another instance of a generalization bound that builds upon the coloring principle; London et al. (2014) later provided another PAC-Bayesian result for dependent inputs. However, one important issue that has not been explored in these studies, is the use of second-order (i.e. variance) information: such information is pivotal to get generalization bounds with fast learning rates as outlined for instance in (Boucheron et al., 2005). To this aim, we here consider the *entropy method* (Boucheron et al., 2003) that is a central technique to obtain concentration inequalities for certain types of functions (namely, sub-additive and self-bounding) and it is at the core of a proof of the well-known Talagrand

concentration inequality for empirical processes (Bousquet, 2002; Ledoux, 1996; Massart, 2000). This inequality makes it possible then to derive generalization bounds based on *Local Rademacher Complexities* (Bartlett et al., 2005; Koltchinskii, 2006) that may induce fast convergence rates. To the best of our knowledge, the question of pairing the entropy method together with the coloring approach has not yet been studied and, we propose to address it in this paper.

Contributions. The main theoretical results of the present paper essentially are of three different kinds. First, we show that, according to the idea of fractional coloring, it is possible to extend the applicability of concentration of certain types of sub-abdditive and self-bounding functions, namely *fractionally sub-additive* and *fractionally self-bounding* functions to the case of dependent variables; the new Bernstein's type concentration inequalities we propose reduce to the usual concentration inequalities when the random variables at hand are independent. Second, thanks to the derived concentration inequality, we introduce the notion of local fractional Rademacher complexity. Finally, we show how these technical results can be instantiated for the learning scenario of bipartite ranking.

Organization of the paper. Section 2 states the general problem we are interested in. In Section 3, we give the formal definition of our framework and explicit the progression of our analysis over the paper. Section 4 presents new entropy-based concentration inequalities which will allow to extend several inequalities proposed for empirical processes to the case of dependent variables, and, in Section 5, we prove a generalization bound for bipartite ranking.

2. Statement of the Problem

Many learning problems deal with interdependent training data. The study of the consistency of the ERM principle requires in this case, the availability of concentration inequalities tailored to handle general functions of dependent random variables. A common example is the reduction of learning problems to classification of pairs of examples like in the bipartite ranking or in multiclass classification with the all-pairs approach (Amini & Usunier, 2015). The former problem deals with the search of a scoring function over a class $\mathcal{F} = \{f : \mathcal{X} \to \mathbb{R}\}$ of real-valued functions using a training set $S = \{(T_i, Y_i)\}_{i=1}^n$ where the observations (T_i, Y_i) are supposed to be identically and independently distributed according to some distribution D, in such a way that $\mathbb{P}_{(T,Y),(T',Y')\sim D}((Y - Y')(f(T) - f(T')) \leq 0)$ is as small as possible. Without loss of generality, we will preferrably term the problem as that of controlling

$$\mathbb{P}_{T^+\sim D_+, T^-\sim D_-}(f(T^+) < f(T^-)), \tag{1}$$

where D_+ (resp. D_-) is the conditional distribution of the positive or relevant (resp. negative or irrelevant) examples. To this end, it is natural to consider some empirical risk $\hat{R}_\ell(f, S)$ on S related to the AUC and defined as

$$\hat{R}_\ell(f, S) \doteq \frac{1}{n_+ n_-} \sum_{i:Y_i=+1} \sum_{j:Y_j=-1} \mathbb{I}_{f(T_i)<f(T_j)}, \tag{2}$$

where $\mathbb{I}_\pi$ is the indicator function that is equal to 1 if the predicate π holds and 0 otherwise. The optimization of (2) can be carried out by finding a classifier of the form $c_f(T, T') = \mathrm{sgn}(f(T) - f(T'))$ that minimizes the classification error over the pairs $(T, +1)$ and $(T', -1)$ (Agarwal & Niyogi, 2005; Clémençon et al., 2008).

Yet, if we consider the random variables $X_{ij} \doteq (T_i^+, T_j^-)$ made of pairs of positive T_i^+ and negative T_j^- examples in S, then each pair X_{ij} is dependent to another pair X_{kl} whenever $i = k$ or $j = l$, and the empirical classification error over these pairs is a function of dependent variables.

In this work, we are interested in deriving concentration inequalities for some functions of dependent variables.

3. Notation and Background Results

3.1. Notation

Throughout, we use the following notation. For any positive integer N, $[N]$ denotes the set $[N] \doteq \{1, \ldots, N\}$. For a sequence $(U_1, \ldots, U_N)$ of elements, that will later refer to sequences of real values or sequence of random variables, and any subset $\mathcal{C}$ of $[N]$, $U_\mathcal{C}$ is the subsequence $U_\mathcal{C} \doteq (U_i)_{i\in\mathcal{C}}$ and, therefore $U_{[N]} = (U_1, \ldots, U_N)$. For $k \in \mathcal{C}$, the sequence $U_\mathcal{C}^{\setminus k}$ is given by $U_\mathcal{C}^{\setminus k} \doteq (U_i)_{i\in\mathcal{C}\setminus\{k\}}$ We assume that $X_{[N]} \doteq (X_1, \ldots, X_N)$ is a sequence of (not necessarily *independent*) random variables taking value in some space $\mathcal{X}$; $\mathcal{A}$ denotes the σ-algebra generated by $X_{[N]}$ and, for $n \in [N]$, $\mathcal{A}_n$ the σ-algebra generated by

$$X_{[N]}^{\setminus n} \doteq (X_1, \ldots, X_{n-1}, X_{n+1}, \ldots, X_N)$$

Further, $f : \mathcal{X}^N \to \mathbb{R}$ is some $\mathcal{A}$-measurable function which allows us to define the random variable Z as:

$$Z \doteq f(X_{[N]})$$

If $X'_{[N]} \doteq (X'_1, \ldots, X'_N)$ is an independent copy of $X_{[N]}$, then, for each n, $X_{[N]}^{(n)}$ and $Z^{(n)}$ are respectively defined as

$$X_{[N]}^{(n)} \doteq (X_1, \ldots, X'_n, \ldots, X_N) \tag{3}$$

$$Z^{(n)} \doteq f(X_{[N]}^{(n)}). \tag{4}$$

Given a subset $\mathcal{C} = \{n_1, \ldots, n_{|\mathcal{C}|}\}$ of $[N]$ and some $n_k \in \mathcal{C}$, $X_\mathcal{C}^{(n_k)}$ is defined as

$$X_\mathcal{C}^{(n_k)} \doteq (X_{n_1}, \ldots, X'_{n_k}, \ldots, X_{n_{|\mathcal{C}|}}) \tag{5}$$

Finally, the expectations taken with respect to $\mathcal{A}_n$ and the σ-algebra genrated by $X_{\mathcal{C}}$ are denoted by $\mathbb{E}_n$ and $\mathbb{E}_{\mathcal{C}}$; when the context is clear, the former is simply denoted by $\mathbb{E}$.

3.2. Concentration of Sub-Additive and Self-Bounding Functions

Essential to some of our results are the notions of sub-additive functions and self-bounding functions.

Definition 1 (Sub-additive functions)**.** A function $f : \mathcal{X}^N \to \mathbb{R}$ of N variables is *sub-additive* if there exists a sequence $(f_n)_{n\in[N]}$ of functions of $N-1$ variables such that for all $x_{[N]} = (x_1, \dots, x_N)$,

$$\sum_{n=1}^{N} (f(x_{[N]}) - f_n(x_{[N]}^{\setminus n})) \leq f(x_{[N]}). \tag{6}$$

Definition 2 (Self-bounding functions)**.** A function $f : \mathcal{X} \to \mathbb{R}$ of N variables is (a, b)-*self-bounding* if there exists a sequence $(f_n)_{n\in[N]}$ of functions of $N-1$ variables such that for all $x_{[N]} = (x_1, \dots, x_N)$,

$$0 \leq f(x_{[N]}) - f_n(x_{[N]}^{\setminus n}) \leq 1, \ \forall n \in [N], \tag{7a}$$

$$\sum_{n=1}^{N} (f(x_{[N]}) - f_n(x_{[N]}^{\setminus n})) \leq a f(x_{[N]}) + b. \tag{7b}$$

The concentration inequalities for sub-additive and self-bounding functions (Boucheron et al., 2009; Bousquet, 2003) are based on bounding the log-Laplace transform function $G : \mathbb{R} \to \mathbb{R}$ defined as

$$G(\lambda) \doteq \log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}Z))]. \tag{8}$$

Using Markov's inequality with the bound exhibited for $G(\lambda)$ together with a clever setting of λ is the traditional way to get concentration inequality *per se*.

As it will shortly appear, it is convenient (and usual) in the setting of concentration results to introduce functions $\psi : \mathbb{R} \to \mathbb{R}$ and $\varphi : [0; +\infty) \to \mathbb{R}$ defined as

$$\psi(x) \doteq \exp(-x) + x - 1 \tag{9}$$

$$\varphi(x) \doteq (1 + x)\log(1 + x) - x. \tag{10}$$

Theorem 1 ((Bousquet, 2002; 2003))**.** *Let $f : \mathcal{X}^N \to \mathbb{R}$ be a sub-additive function. Assume $X_{[N]}$ is a sequence of* independent *random variables and that $(Y_n)_{n\in[N]}$ is a sequence of real-valued $\mathcal{A}$-measurable random variables such that $\mathbb{P}(Y_n \leq Z - Z^{(n)} \leq 1) = 1$ and $\mathbb{P}(\mathbb{E}_n[Y_n] \geq 0) = 1$. Let $\sigma^2 \in \mathbb{R}$ be such that $\mathbb{P}\left(\sigma^2 \geq \sum_{n=1}^{N} \mathbb{E}_n\left[Y_n^2\right]\right) = 1$, If there exists $b > 0$ such that for all $n \in [N]$, $\mathbb{P}(Y_n \leq b) = 1$ then, for all $\lambda \geq 0$*

$$G(\lambda) \leq \psi(-\lambda)v \tag{11}$$

where $v \doteq (1 + b)\mathbb{E}Z + \sigma^2$.

We recall the following result which provides a bound on the expectation of the Laplace transform of self-bounding functions due to McDiarmid & Reed (2006). Note that similar results for *weakly self-bounding*—a slightly weaker notion than self-bounding— functions are given by Maurer (2006) and this set of results were refined (with better constants) by Boucheron et al. (2009). We decide to refer to the result of McDiarmid & Reed (2006) because it implies upper and lower tail bound in a more compact way—but the extensions to the dependent case that we propose also apply to the more recent versions of the results.

Theorem 2 ((McDiarmid & Reed, 2006))**.** *Let $f : \mathcal{X}^N \to \mathbb{R}$ be a (a, b)-self-bounding function. Assume $X_{[N]}$ is a sequence of* independent *random variables. Let $\mu \doteq \mathbb{E}Z$. The following holds.*

$$G(\lambda) \leq (a\mu + b)\psi(\lambda), \qquad \forall \lambda \leq 0 \tag{12}$$

$$G(\lambda) \leq \frac{a\mu + b}{2(1 - a\lambda)}\lambda^2, \qquad 0 \leq \lambda \leq 1/a. \tag{13}$$

3.3. Dependency Graph

Specific graphs, namely dependency graphs, are also at the core of the present study: a graph $G = (V, E)$ is made of a finite set V of vertices and a set $E \subseteq V \times V$ of edges that connect the vertices. We have the following definition of an exact proper fractional cover of a graph that we will make intensive use of afterwards.

Definition 3 (Exact proper fractional cover of G)**.** Let $G = (V, E)$ be a graph. $\mathcal{C} = \{(\mathcal{C}_j, \omega_j)\}_{j\in[J]}$, for some positive integer J, with $\mathcal{C}_j \subseteq V$ and $\omega_j \in [0, 1]$ is an exact proper fractional cover of G, if:

1. it is *proper:* $\forall j$, $\mathcal{C}_j$ is an *independent set*, i.e., there is no connections between vertices in $\mathcal{C}_j$;
2. it is an *exact fractional cover* of G: $\forall v \in V$, $\sum_{j:v\in\mathcal{C}_j} \omega_j = 1$.

The weight $W(\mathcal{C})$ of $\mathcal{C}$ is given by: $W(\mathcal{C}) \doteq \sum_{j\in[J]} \omega_j$ and the minimum weight $\chi^*(G) = \min_{\mathcal{C}\in\mathcal{K}(G)} W(\mathcal{C})$ over the set $\mathcal{K}(G)$ of all exact proper fractional covers of G is the *fractional chromatic number* of G.

Note that, as observed by Janson (2004), Lemma 3.2, we may restrict ourselves to working with *exact* fractional covers, which requires $\forall v \in V$, $\sum_{j:v\in\mathcal{C}_j} \omega_j = 1$ instead of the weaker condition $\forall v \in V$, $\sum_{j:v\in\mathcal{C}_j} \omega_j \geq 1$ for non-exact fractional covers, without loss of generality, since any fractional covers induces an exact fractional cover.

From now on, it must be understood that we refer to exact proper fractional cover when using the simpler term of fractional cover. The reader that is not familiar with this notion of fractional covers may regard them as generalization of graph coloring, where the question is to assign the smallest

number of colors to nodes of a graph so that no two connected nodes share the same color. When colored this way, the set of points that have the same color are necessarily independent sets and the coloring might be thought of an exact proper fractional coloring with every $\mathcal{C}_j$ corresponding to color j and every ω_j being equal to 1. Given some graph G, the smallest number of colors $\chi(G)$ is its *chromatic number* and the following holds: $\chi_f(G) \leq \chi(G)$.

Definition 4 (Dependency Graph)**.** Let $X_{[N]} \doteq (X_1, \ldots, X_N)$ be a sequence of random variables. We may associate the dependency graph $G_X \doteq (V_X, E_X)$ to $X_{[N]}$ so that **i)** $V_X = [N]$ and **ii)** $(i,j) \in E_X$ if and only if X_i and X_j are dependent random variables.

Note that there are other notions of dependency graphs that can be envisioned (see (Janson, 2004)). The present notion of dependency graph will however suffice to our purpose. As we shall see, computing an exact proper fractional cover $\{(\mathcal{C}_j, \omega_j)\}_{j\in[J]}$ of G_X allows one to decompose $X_{[N]}$ in sets of independent variables $X_{\mathcal{C}_j}$. This will make it possible to have the usual concentration inequalities for independent variables to carry over to the dependent case.

4. New Concentration Inequalities

As stated above, we build upon the works on the entropy method (see, for a somewhat exhaustive overview the method the work of Boucheron et al. (2003)) and that of Janson (2004) to provide new concentration inequalities for functions of dependent random variables.

4.1. Fractionally Colorable Functions

We aim at establishing concentration results for functions that are more complex than sums of dependent random variables. To this end, we introduce the notions of *fractionally colorable functions*, *colorable sub-additive functions* and *colorable self-bounding functions*; they refine the definitions given in the previous section.

Definition 5 (Fractionally colorable function)**.** Let $G = ([N], E)$ be a graph. A function $f : \mathcal{X}^N \to \mathbb{R}$ is *fractionally colorable* with respect to G if there exists a *decomposition* $\mathcal{D}_G(f) = \{(f_j, \mathcal{C}_j, \omega_j)\}_{j\in[J]}$ of J triplets, such that:

1. $\mathcal{C} = \{(\mathcal{C}_j, \omega_j)\}_{j\in[J]}$ is an exact proper fractional cover of G;
2. for all j, $f_j : \mathcal{X}^{|\mathcal{C}_j|} \to \mathbb{R}$ is a function of $|\mathcal{C}_j|$ variables and f decomposes as

$$\forall x_{[N]} \in \mathcal{X}^N,\ f(x_{[N]}) = \sum_j \omega_j f_j(x_{\mathcal{C}_j}) \qquad (14)$$

The decomposition $\mathcal{D}_G(f)$ of f is *optimal* if the weight of the cover $\mathcal{C} = \{(\omega_j, \mathcal{C}_j)\}_{j\in[J]}$ is the smallest over all decompositions of f. In that case, the *chromatic decomposition number* χ_f of f is the weight of such an optimal decomposition.

In the sequel, and without loss of generality, we will always consider optimal decompositions of fractionally colorable functions. Also, we will assume that the graph G at hand is the dependency graph of the sequence $X_{[N]}$ under study.

We may now assume that we are working with a fractionally colorable function f and we may recall/introduce notation: as before, Z and $Z^{(n)}$ are defined as

$$Z \doteq f\left(X_{[N]}\right), \qquad Z^{(n)} \doteq f\left(X^{(n)}_{[N]}\right),$$

and, for $j \in [J], n \in \mathcal{C}_j$, Z_j and $Z_j^{(n)}$ are defined as:

$$Z_j \doteq f_j\left(X_{\mathcal{C}_j}\right), \qquad Z_j^{(n)} \doteq f_j\left(X^{(n)}_{\mathcal{C}_j}\right),$$

where $X^{(n)}_{\mathcal{C}_j}$ is defined as in equation (5). Hence,

$$Z = \sum_j \omega_j Z_j. \qquad (15)$$

Let Π_J be the family of discrete probability distributions over J-sets:

$$\Pi_J \doteq \left\{(p_1, \ldots, p_J) : \sum_{j=1}^J p_j = 1 \text{ and } p_j > 0, \forall j\right\} \qquad (16)$$

We then have the following central lemma.

Lemma 1 (Central Lemma)**.** *If f is fractionally colorable then $\forall (p_1, \ldots, p_J) \in \Pi_J, \forall \lambda \in \mathbb{R}$,*

$$G(\lambda) \leq \log \sum_{j\in[J]} p_j \exp\left[G_j\left(\lambda \frac{\omega_j}{p_j}\right)\right] \qquad (17)$$

where

$$G_j(\lambda) \doteq \log \mathbb{E}_{\mathcal{C}_j}\left[\exp\left(\lambda\left(Z_j - \mathbb{E}_{\mathcal{C}_j}[Z_j]\right)\right)\right]. \qquad (18)$$

Proof.

$$\begin{aligned}
G(\lambda) &= \log \mathbb{E}[\exp(\lambda(Z - \mathbb{E}[Z]))] \\
&= \log \mathbb{E}\left[\exp\left(\lambda \sum_{j\in[J]} \omega_j \left(Z_j - \mathbb{E}_{\mathcal{C}_j}[Z_j]\right)\right)\right] \\
&= \log \mathbb{E}\left[\exp\left(\lambda \sum_{j\in[J]} p_j \frac{\omega_j}{p_j} \left(Z_j - \mathbb{E}_{\mathcal{C}_j}[Z_j]\right)\right)\right] \\
&\leq \log \mathbb{E}\left[\sum_{j\in[J]} p_j \exp\left(\lambda \frac{\omega_j}{p_j} \left(Z_j - \mathbb{E}_{\mathcal{C}_j}[Z_j]\right)\right)\right] \\
&\qquad \text{(convexity of } x \mapsto e^x \text{ and the Jensen inequality)} \\
&= \log \sum_{j\in[J]} p_j \mathbb{E}_{\mathcal{C}_j}\left[\exp\left(\lambda \frac{\omega_j}{p_j} \left(Z_j - \mathbb{E}_{\mathcal{C}_j}[Z_j]\right)\right)\right]
\end{aligned}$$

□

Remark that the functions G_j's are the counterparts of G for the random variables Z_j, which are defined with respect to the set $\mathcal{C}_j$ of *independent* variables.

4.2. Concentration of Fractionally Sub-Additive Functions

Definition 6 (Fractionally Sub-Additive function)**.** Let $G = ([N], E)$ be some graph. A function $f : \mathcal{X}^N \to \mathbb{R}$ is *fractionally sub-additive* if it is fractionally colorable with respect to G with decomposition $\mathcal{D}_G(f) = \{(f_j, \mathcal{C}_j, \omega_j)\}_{j\in[J]}$ and each f_j is sub-additive.

Proposition 1. *Suppose the following assumptions are true. f is fractionally sub-additive with decomposition $\mathcal{D}(f) = \{(f_j, \mathcal{C}_j, \omega_j)\}_{j\in[J]}$. Assume that for all $j \in [J]$:*

- *$(Y_{j,n})_{n\in\mathcal{C}_j}$ is a sequence of real-valued $\sigma(X_{\mathcal{C}_j})$-measurable random variables such that $\forall n \in \mathcal{C}_j$,*

$$\mathbb{P}(Y_{j,n} \leq Z_j - Z_j^{(n)} \leq 1) = 1,$$
$$\mathbb{P}(\mathbb{E}_{j,n}[Y_{j,n}] \geq 0) = 1,$$

  where $\mathbb{E}_{j,n}$ denotes the expectation with respect to the σ-algebra generated by $(X_{\mathcal{C}_j\setminus\{n\}})$;
- *there exists $\sigma_j^2 \in \mathbb{R}$ so that*

$$\mathbb{P}\left(\sigma_j^2 \geq \sum_{n\in\mathcal{C}_j} \mathbb{E}_{j,n}\left[Y_{j,n}^2\right]\right) = 1;$$

- *there exists a positive $b_j \in \mathbb{R}$ such that $\forall n \in \mathcal{C}_j$, $\mathbb{P}(Y_{j,n} \leq b_j) = 1$;*
- *$v_j \in \mathbb{R}$ denotes the real $v_j \doteq (1 + b_j)\mathbb{E}[Z_j] + \sigma_j^2$.*

The following result holds: for all $\lambda \geq 0$ and for all $(p_1, \ldots, p_J) \in \Pi_J$,

$$G(\lambda) \leq \log \sum_j p_j \exp\left(v_j \psi\left(-\lambda \frac{\omega_j}{p_j}\right)\right), \tag{19}$$

where ψ is defined as in (9).

Proof. Let $(p_1, \ldots, p_J) \in \Pi_J$ and $\lambda > 0$ (the proof is trivially true for $\lambda = 0$).

$$\begin{aligned} G(\lambda) &\leq \log \sum_{j\in[J]} p_j \exp\left[G_j\left(\lambda \frac{\omega_j}{p_j}\right)\right] && \text{(Lemma 1)} \\ &\leq \log \sum_{j\in[J]} p_j \exp\left(v_j \psi\left(-\lambda \frac{\omega_j}{p_j}\right)\right) && \text{(Theorem 1)} \end{aligned}$$

□

The following result extends Bennett's inequality presented by (Bousquet, 2002) to the dependent case.

Theorem 3 (Bennett's Inequality for Dependent Variables)**.** *Suppose the assumptions of Proposition 1 hold with $b_1 = \ldots = b_J \doteq b$ and define the constants*

$$\sigma^2 \doteq \sum_{j\in[J]} \omega_j \sigma_j^2,\ v \doteq (1 + b)\mathbb{E}[Z] + \sigma^2,\ c \doteq 25\chi_f/16.$$

The following results hold:

- *for all $t \geq 0$*

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-\frac{v}{\chi_f}\varphi\left(\frac{4t}{5v}\right)\right); \tag{20}$$

  where φ is defined as in (10).
- *for all $t \geq 0$*

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2cvt} + \frac{ct}{3}\right) \leq e^{-t}. \tag{21}$$

Proof. To get the results, we start from Proposition 1 and we follow the steps of Janson (2004) (Theorem 3.4). Set

$$v \doteq \sum_{j\in[J]} \omega_j v_j,\ W \doteq \sum_{j\in[J]} \omega_j, \tag{22}$$
$$U \doteq \sum_{j\in[J]} \omega_j \max\left(1, v_j^{1/2} W^{1/2} v^{-1/2}\right), \tag{23}$$
$$p_j \doteq \omega_j \max\left(1, v_j^{1/2} W^{1/2} v^{-1/2}\right)/U. \tag{24}$$

With these choices, we observe that each summand of the sum in (19) is such that $v_j\psi(-\lambda U\omega_j/p_j) \leq v\psi(-\lambda U)/W$. Indeed, if $v_j^{1/2}W^{1/2}v^{-1/2} \leq 1$, then $p_j = \omega_j/U$, $v_j \leq v/W$, and

$$v_j\psi\left(-\lambda\frac{\omega_j}{p_j}\right) = v_j\psi(\lambda U) \leq \frac{v}{W}\psi(-\lambda U).$$

Otherwise (i.e. $v_j^{1/2}W^{1/2}v^{-1/2} > 1$)

$$p_j = \omega_j v_j^{1/2} W^{1/2} v^{-1/2}/U$$

and

$$\begin{aligned} v_j\psi\left(-\lambda\frac{\omega_j}{p_j}\right) &= v_j\psi\left(-\lambda U\frac{v^{1/2}}{v_j^{1/2}W^{1/2}}\right) \\ &\leq v_j\left(\frac{1}{W^{1/2}}\right)^2\psi(-\lambda U) = \frac{v_j}{W}\psi(-\lambda U), \end{aligned}$$

where the inequality comes from a property of ψ that is recalled in Proposition 5 (in Appendix A).

This bounding of $v_j\psi(-\lambda U\omega_j/p_j)$, and the fact that $x \mapsto e^x$ is an increasing function give

$$\sum_{j\in[J]} p_j \exp\left(v_j\psi\left(-\lambda U\frac{\omega_j}{p_j}\right)\right) \leq \exp\left(\frac{v}{W}\psi(-\lambda U)\right).$$

Using Markov's inequality (cf. Theorem 6, Appendix A),

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) = \mathbb{P}\left(\exp\left(\lambda(Z - \mathbb{E}[Z])\right) \geq \exp(\lambda t)\right) \leq \exp\left(\frac{v}{W}\psi(-\lambda U) - \lambda t\right).$$

The upper bound of this inequality is minimized for $\lambda = \ln(1 + tW/vU)/U$; plugging in this value yields

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{v}{W}\varphi\left(\frac{tW}{Uv}\right)\right).$$

Now using the fact that $\forall x \in \mathbb{R}, x \leq 1 + x^2/4$, we get

$$U \leq \sum_j \omega_j \left(1 + \frac{v_j W}{4v}\right) = W + \frac{vW}{4v} = \frac{5}{4}W.$$

Since $x \mapsto \varphi(t/x)$ is decreasing for $t > 0$, we readily have the following upper bound

$$\mathbb{P}(Z - \mathbb{E}[Z] \geq t) \leq \exp\left(-\frac{v}{W}\varphi\left(\frac{4t}{5v}\right)\right).$$

As said before, we consider only optimal decompositions of f and the total weight W may be readily replaced by the chromatic number χ_f. Finally, we observe that:

$$v = \sum_j \omega_j v_j = \sum_j \omega_j((1+b)\mathbb{E}[Z_j] + \sigma_j^2) = (1+b)\mathbb{E}[Z] + \sigma^2.$$

Inequality (21) is deduced from (20) and the fact that $x \geq 0, \varphi(x) \geq x^2/(2(1 + x/3)$. □

This, in turn, gives the following Talagrand's type inequality for empirical processes in the dependent case.

Theorem 4. *Let $\mathcal{F}$ be a set of functions from $\mathcal{X}$ to $\mathbb{R}$ and assume all functions in $\mathcal{F}$ are measurable, square-integrable and satisfy $\mathbb{E}[f(X_n)] = 0, \forall n \in [N]$ and $\sup_{f\in\mathcal{F}} \|f\|_\infty \leq 1$. Assume that $\mathcal{C} = \{(\mathcal{C}_j, \omega_j)\}$ is a cover of the dependency graph of $X_{[N]}$ and let $\chi_f \doteq \sum_j \omega_j$.*

Let us define:

$$Z \doteq \sum_{j\in[J]} \omega_j \sup_{f\in\mathcal{F}} \sum_{n\in\mathcal{C}_j} f(X_n)$$

Let σ_j be so that $\sigma_j^2 \geq \sum_{n\in\mathcal{C}_j} \sup_{f\in\mathcal{F}} \mathbb{E}\left[f^2(X_n)\right]$.

Let $v \doteq \sum_j \omega_j \sigma_j^2 + 2\mathbb{E}[Z]$. For any $t \geq 0$,

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-\frac{v}{\chi_f}\varphi\left(\frac{4t}{5v}\right)\right) \tag{25}$$

Also, if $c \doteq 25\chi_f/16$.

$$\mathbb{P}\left(Z \geq \mathbb{E}[Z] + \sqrt{2cvt} + \frac{ct}{3}\right) \leq e^{-t} \tag{26}$$

Proof. (Sketch.) The proof is similar to the one of Bousquet (2003) (Theorem 7.3) and it hinges on the fact that, by Lemma 2, Appendix A, the Z_j's are indeed sub-additive functions and by studying the random variables $Y_{j,n}$ defined for $n \in \mathcal{C}_j$ as: $Y_{j,n} \doteq f_j^n(X_n)$, where f_n^j is such that $\sum_{k\in\mathcal{C}_j\setminus\{n\}} f_j^n(X_k) = \sup_{f\in\mathcal{F}} \sum_{k\in\mathcal{C}_j} f(X_k)$ which also yields that $b = 1$ in the definition of v. □

We may now introduce the *Local Fractional Rademacher Complexity* which, combined with the previous inequality, is useful to get generalization bounds (see Section 5).

Definition 7. The *Local Fractional Rademacher Complexity* $\mathcal{R}(\mathcal{F}, r)$ is defined as

$$\mathcal{R}(\mathcal{F}, r) \doteq \frac{2}{N}\mathbb{E}_\xi \sum_{j\in[J]} \omega_j \mathbb{E}_{X_{\mathcal{C}_j}} \sup_{f\in\mathcal{F}:\mathbb{V}f\leq r} \sum_{n\in\mathcal{C}_j} \xi_i f(X_n) \tag{27}$$

where $\xi = (\xi_1, \ldots, \xi_N)$ is a sequence of N independent Rademacher variables: $\mathbb{P}(\xi_n = 1) = \mathbb{P}(\xi_n = -1) = 1/2$.

This is a generalization of the fractional Rademacher complexity of Usunier et al. (2006). The following holds:

Proposition 2. *For all $r > 0$,*

$$\mathbb{E}_{X_{[N]}} \sum_j \omega_j \sup_{f\in\mathcal{F}:\mathbb{V}f\leq r} \sum_{n\in\mathcal{C}_j} [\mathbb{E}f(X_n) - f(X_n)] \leq N\mathcal{R}(\mathcal{F}, r).$$

Proof. A simple symmetrization argument carefully used in combination with the fractional decomposition of f gives the result. □

4.3. Concentration of Fractionally Self-Bounding Functions

We provide concentration inequalities for a generalization of self-bounding functions, namely, fractionally self-bounding functions. Such results may have some use to problems that naturally make self-bounding functions appear (Boucheron et al., 2009; McDiarmid & Reed, 2006).

Definition 8 (Fractionally Self-Bounding Function)**.** Let $G = ([N], E)$ be some graph. A function $f : \mathcal{X}^N \to \mathbb{R}$ with decomposition $\mathcal{D}_G(f) = \{(f_j, \mathcal{C}_j, \omega_j)\}_{j\in[J]}$ is $(\{a_j\}_j, \{b_j\}_j)$-*fractionally self-bounding* if each f_j is (a_j, b_j)-self-bounding.

Proposition 3. *Let G_X be the dependency graph associated with $X_{[N]}$. Let $f : \mathcal{X}^N \to \mathbb{R}$ be $(\{a_j\}_j, \{b_j\}_j)$-fractionally self-bounding. The following holds for all $(p_1, \ldots, p_J) \in \Pi_J$,*

- *for all $\lambda \leq 0$*

$$G(\lambda) \leq \log \sum_{j\in[J]} p_j \exp\left((a_j\mu_j + b_j)\psi\left(\lambda\frac{\omega_j}{p_j}\right)\right) \tag{28}$$

- *for all* $0 \leq \lambda < \min(p_j/(a_j\omega_j))$, *with* $\mu_j = \mathbb{E}_{\mathcal{C}_j}[Z_j]$,

$$G(\lambda) \leq \log \sum_{j\in[J]} p_j \exp\left(\frac{a_j\mu_j + b_j}{2(1-\lambda a_j\omega_j/p_j)}\left(\frac{\lambda\omega_j}{p_j}\right)^2\right). \tag{29}$$

Proof. A combination of Lemma 1 and Theorem 2. □

This proposition entails the following concentration inequalities for the upper and lower tails of Z.

Theorem 5. *Let* $f : \mathcal{X}^N \to \mathbb{R}$ *be* $(\{a_j\}_j, \{b_j\}_j)$*-fractionally self-bounding, with decomposition* $\mathcal{D}_{G_X}(f) = \{(f_j, \mathcal{C}_j, \omega_j)\}_{j\in[J]}$. *Define*

$$\gamma_j \doteq (a_j\mu_j + b_j),\ \gamma \doteq \sum_{j\in[J]} \omega_j\gamma_j,\ \chi_f \doteq \sum_j \omega_j.$$

The following results hold: for all $t > 0$,

$$\mathbb{P}(Z \leq \mathbb{E}[Z] - t) \leq \exp\left(-\frac{\gamma}{\chi_f}\varphi\left(\frac{4t}{5\gamma}\right)\right), \tag{30}$$

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-\frac{\gamma}{\chi_f}\varphi\left(\frac{t}{\gamma\rho}\right)\right), \tag{31}$$

where $\rho \doteq \frac{1}{\chi_f}\sum_{j\in[J]}\omega_j a_j + \frac{1}{4\gamma}\sum_{j\in[J]}\omega_j\frac{\gamma_j}{a_j}$.

If $a \doteq a_1 = \ldots = a_J$, *then* ρ *simplifies to* $\rho = a + \frac{1}{4a}$ *and* (31) *takes the more convenient form*

$$\mathbb{P}(Z \geq \mathbb{E}[Z] + t) \leq \exp\left(-\frac{\gamma}{\chi_f}\varphi\left(\frac{4at}{(4a^2+1)\gamma}\right)\right), \tag{32}$$

which is similar to Equation (30) *for* $a = 1$.

Proof. The proof of Equation (30) follows exactly the same steps as the proof of Theorem 3 after noticing that the starting point of the former, namely Equation (28), is similar to the Equation (19), the starting point of the latter, with λ in place of $-\lambda$ and γ_j in place of v_j.

The proof of Equation (31) hinges on the following choices:

$$U \doteq \sum_{j\in[J]} a_j\omega_j \max\left(1, a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2}\right),$$
$$p_j \doteq a_j\omega_j \max\left(1, a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2}\right)/U,$$

which allow the argument of the exponential in the right-hand side of Equation (29) to be rewritten as, for all $j \in [J]$

$$\theta_j(\lambda) \doteq \frac{\gamma_j}{2(1-\lambda a_j\omega_j/p_j)}\left(\frac{\lambda\omega_j}{p_j}\right)^2 = \frac{\gamma_j/a_j^2 \cdot U^2/\max^2\left(1, a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2}\right)}{2\left(1-\lambda U/\max\left(1, a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2}\right)\right)}\lambda^2.$$

It is easy to verify that for all $\lambda \geq 0$ we have

$$\theta_j(\lambda) \leq \frac{\gamma/\chi_f}{2\,(1-\lambda U)}(\lambda U)^2$$

Indeed, if $a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2} \leq 1$ then θ_j is bounded as

$$\theta_j(\lambda) = \frac{\gamma_j/a_j^2}{2\,(1-\lambda U)}(\lambda U)^2 \leq \frac{\gamma/\chi_f}{2\,(1-\lambda U)}(\lambda U)^2.$$

And if, $a_j^{-1}\gamma_j^{1/2}\chi_f^{1/2}\gamma^{-1/2} > 1$, then

$$\begin{aligned}\theta_j(\lambda) &= \frac{\gamma/\chi_f}{2\left(1-\lambda U\cdot a_j\gamma_j^{-1/2}\chi_f^{-1/2}\gamma^{1/2}\right)}(\lambda U)^2 \\ &\leq \frac{\gamma/\chi_f}{2\left(1-\lambda U\cdot \chi_f^{1/2}\gamma^{-1/2}\chi_f^{-1/2}\gamma^{1/2}\right)}(\lambda U)^2 \\ &= \frac{\gamma/\chi_f}{2\,(1-\lambda U)}(\lambda U)^2,\end{aligned}$$

where the upper-bounding is possible because $\lambda \geq 0$. Therefore, from the bound (29) it comes

$$\exp G(\lambda) = \mathbb{E}\left[\exp\left(\lambda(Z - \mathbb{E}[Z])\right)\right] \leq \exp\left(\frac{\gamma/\chi_f}{2\,(1-\lambda U)}(\lambda U)^2\right).$$

Using Markov's inequality, and the fact that $p_j/(a_j\omega_j) \geq 1/U$ for all $j \in [J]$, we get

$$\begin{aligned}\mathbb{P}(Z - \mathbb{E}[Z] \geq t) &= \mathbb{P}(\exp\left(\lambda(Z-\mathbb{E}[Z])\right) \geq \exp(\lambda t)) \\ &\leq \exp\left(\inf_{0\leq\lambda<1/U}\frac{\gamma/\chi_f}{2\,(1-\lambda U)}(\lambda U)^2 - \lambda t\right) \\ &= \exp\left(-\frac{\gamma}{\chi_f}\varphi\left(\frac{\chi_f t}{\gamma U}\right)\right),\end{aligned}$$

where we used Lemma 3 (Appendix A) to get the last line. Using the inequality $\forall x \in \mathbb{R}, x \leq 1 + x^2/4$ once again, we may bound U as

$$U \leq \sum_{j\in[J]} a_j\omega_j\left(1 + \frac{\chi_f\gamma_j}{4a_j^2\gamma}\right) = \sum_{j\in[J]}\omega_j a_j + \frac{\chi_f}{4\gamma}\sum_{j\in[J]}\omega_j\frac{\gamma_j}{a_j}$$

and use the fact that $x \mapsto \varphi(t/x)$ is decreasing for $t > 0$. □

5. Induced Generalization Bounds for Bipartite Ranking

In this section, we show how the concentration inequalities we established in the previous section can be of some use to derive generalization bounds for predictors trained on interdependent data. We will more precisely take advantage of the concentration inequality given by Theorem 3 and provide a generalization bound for the problem of bipartite ranking that can be reduced to the classification of pairs of examples (see Section 2). To do so, our proof will

rest on the notion of local fractional Rademacher complexity, a generalization of both notions of local Rademacher complexity (Bartlett et al., 2005; Koltchinskii, 2006) and fractional Rademacher complexity (Usunier et al., 2006). If we assume that $n_+ \leq n_-$, it is easy to see that $\mathcal{C} \doteq \{(\mathcal{C}_j, \omega_j \doteq 1)\}_{j\in[n^-]}$ with

$$\mathcal{C}_j \doteq \{(i, (j+i-2 \mod n_-)+1) : i = 1, \ldots, n_+\}$$

is an exact cover of the dependency graph of $(X_{ij})_{ij}$. The chromatic number of this cover is therefore $\chi_f = \sum_j \omega_j = n_-$ and $\hat{R}_\ell(f, S)$ decomposes as

$$\hat{R}_\ell(f, S) = \frac{1}{N} \sum_{j\in[n_-]} \sum_{(k,l)\in\mathcal{C}_j} \ell(f, X_{kl}), \tag{33}$$

where, abusing notation, $\ell(f, X_{kl}) = \ell(f, T_i^+, T_j^-)$ and $N \doteq n_+ n_-$. This is a colorable function with respect to the sequence $\underline{X} \doteq (X_{kl})_{kl}$ and the tools we have developed will help us derive a generalization bound for f.

Given a family of functions $\mathcal{F}$, and $r > 0$, we define the parameterized family $\mathcal{F}_{\ell,r}$ which, for $r > 0$, is given by

$$\mathcal{F}_{\ell,r} \doteq \{f : f \in \mathcal{F}, \mathbb{V}_{X_{1,1}} \ell(f, X_{1,1}) \leq r\},$$

where $\mathbb{V}$ denotes the variance (recall that all the X_{kl} are identically distributed). Now denote the function Φ as

$$\Phi(\underline{X}, r) \doteq N \sup_{f\in\mathcal{F}_{\ell,r}} \left[\mathbb{E}_{\underline{X}'}[\hat{R}_\ell(f, \underline{X}')] - \hat{R}_\ell(f, \underline{X})\right],$$

where $\underline{X}'$ is a copy $\underline{X}$ and where we have used the notation $\mathbb{E}_{\underline{X}'}[\hat{R}_\ell(f, \underline{X}')]$ for $\mathbb{E}_S \hat{R}_\ell(f, S)$ to make explicit the dependence on the sequence of dependent variables $\underline{X}'$. It is easy to see that

$$\Phi(\underline{X}, r) \leq \sum_{j\in[n_-]} \sup_{f\in\mathcal{F}_{\ell,r}} \sum_{(k,l)\in\mathcal{C}_j} \left[\mathbb{E}_{X'_{kl}}[\ell(f, X'_{kl})] - \ell(f, X_{kl})\right] \doteq Z. \tag{34}$$

When ℓ takes values in the interval $[0, 1]$ then Theorem 4 readily applies to upper bound the right hand side of (34). Therefore, for $t > 0$, the following holds with probability at least $1 - e^{-t}$:

$$\Phi(\underline{X}, r) \leq \mathbb{E}[Z] + \sqrt{2cvt} + \frac{ct}{3}$$

where $c = 25\chi_f/16 = 25n_-/16$ and $v \leq Nr + 2\mathbb{E}[Z]$. Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ and $2\sqrt{ab} \leq ua + b/u$ for all $u > 0$, we further have, for all $\alpha > 0$

$$\Phi(\underline{X}, r) \leq (1+\alpha)\mathbb{E}[Z] + \sqrt{2cNrt} + \left(\frac{1}{3} + \frac{1}{\alpha}\right) ct,$$

and, using Proposition 2, we get, the following proposition.

Proposition 4. *With probability $1 - e^{-t}$, for all $\alpha > 0$*

$$\Phi(\underline{X}, r) \leq (1+\alpha) N \mathcal{R}(\mathcal{F}_{\ell,r}, r) + \sqrt{2cNrt} + \left(\frac{1}{3} + \frac{1}{\alpha}\right) ct,$$

or, using $\chi_f = n_-$, $N = n_+ n_-$, with probability at least $1 - e^{-t}$, for all $f \in \mathcal{F}_{\ell,r}$

$$\mathbb{E}_S[\hat{R}(f, S)] - \hat{R}(f, S) \leq \inf_{\alpha>0} \left((1+\alpha)\mathcal{R}(\mathcal{F}_{\ell,r}, r) + \frac{5}{4}\sqrt{\frac{2rt}{n_+}} + \frac{25}{16}\left(\frac{1}{3} + \frac{1}{\alpha}\right)\frac{t}{n_+}\right).$$

As is common with generalization bounds for bipartite ranking, the convergence rate is governed by the least represented class, i.e. the positive class here. Note this result is only the starting point of a wealth of results that may be obtained using the concentration inequalities studied here. In particular, it might be possible to sutdy how arguments based on star hulls and subroot functions may help us to get fast-rate-like results akin to (Clémençon et al., 2008).

6. Conclusion

We have proposed new concentration inequalities for functions of dependent variables. From these, we derived a new Talagrand concentration inequality for fractionally sub-additive functions and fractionally self-bounding functions of dependent variables. An instance of a generalization bounds based on Fractional Local Rademacher Complexity for bipartite ranking exemplifies the usefulness of our concentration results.

Acknowledgments. This work is partially supported by the French GIP ANR under contract ANR GRETA 12-BS02-004-01 Greediness: theory and algorithms, and the LabEx PERSYVAL- Lab ANR-11-LABX-0025.

A. Technical Results

Theorem 6 (Markov Inequality)**.** *Let X be a nonnegative random variable. For all $a > 0$ $\mathbb{P}(X > a) \leq \frac{\mathbb{E}[X]}{a}$.*

Proposition 5 (Lemma A.3 of Bousquet (2003))**.** *Let g_λ be as $g_\lambda \doteq \psi(-\lambda x)/x^2$. If $\lambda \geq 0$ then g_λ is non-decreasing on $\mathbb{R}$. If $\lambda \leq 0$ then g_λ is non-increasing on $\mathbb{R}$.*

Lemma 2 (Lemma C.1 of Bousquet (2003))**.** *Let $\mathcal{F}$ be a set of functions and let $Z \doteq \sup_{f\in\mathcal{F}} \sum_{k=1}^n f(X_k)$. Then, defining $Z_k \doteq \sup_{f\in\mathcal{F}} \sum_{i\neq k} f(X_i)$, Z is sub-additive. The same is true if $Z \doteq \sup_{f\in\mathcal{F}} |\sum_{k=1}^n f(X_k)|$ and $Z_k \doteq \sup_{f\in\mathcal{F}} \left|\sum_{i\neq k} f(X_i)\right|$.*

Lemma 3 (Lemma 11 of Boucheron et al. (2003))**.** *Let C and a denote two positive real numbers. Then*

$$sup_{\lambda\in[0,1/a)} \left(\lambda t - \frac{C\lambda^2}{1-a\lambda}\right) = \frac{2C}{a^2}\varphi\left(\frac{at}{2C}\right),$$

and the supremum is at $\lambda = \frac{1}{a}\left(1 - \left(1 + \frac{at}{C}\right)^{-1/2}\right)$.

References

Agarwal, S. and Niyogi, P. Stability and Generalization of Bipartite Ranking Algorithms. In *COLT*, pp. 32–47, 2005.

Amini, M.-R. and Usunier, N. *Learning with Partially Labeled and Interdependent Data*. Springer, 2015.

Bartlett, P., Bousquet, O., and Mendelson, S. Local Rademacher complexities. *Annals of Statistics*, 33(4): 1497–1537, 2005.

Boucheron, S., Lugosi, G., and Massart, P. Concentration inequalities using the entropy method. *The Annals of Probability*, 31(3):1583–1614, 2003.

Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification : A survey of some recent advances. *ESAIM. P&S*, 9:323–375, 2005.

Boucheron, S., Lugosi, G., and Massart, P. On concentration of self-bounding functions. *Electronic Journal of Probability*, 14(64):1884–1899, 2009.

Bousquet, O. A Bennett Concentration Inequality and its Application to Suprema of Empirical Processes. *CRAS, Serie I*, 334:495–500, 2002.

Bousquet, O. Concentration Inequalities for Sub-Additive Functions Using the Entropy Method. In Giné, E., Houdré, C., and Nualart, D. (eds.), *Stochastic Inequalities and Applications*, volume 56 of *Progress in Probability*, pp. 213–247. Birkhäuser Basel, 2003.

Clémençon, S., Lugosi, G., and Vayatis, N. Ranking and empirical minimization of u -statistics. *The Annals of Statistics*, 36(2):844–874, April 2008. ISSN 0090-5364.

Janson, S. Large Deviations for Sums of Partly Dependent Random Variables. *Random Structures and Algorithms*, 24(3):234–248, 2004.

Karandikar, R. L. and Vidyasagar, M. Rates of uniform convergence of empirical means with mixing processes. *Statistics and Probability Letters*, 58(3):297307, 2002.

Koltchinskii, V. Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 12 2006.

Kontorovich, L. and Ramanan, K. Concentration inequalities for dependent random variables via the martingale method. *The Annals of Probability*, 36(6):2126–2158, 2008.

Ledoux, M. Talagrand deviation inequalities for product measures. *ESAIM: Probabilty and Statistics*, 1:63–87, 1996.

London, B., Huang, B., Taskar, B., and Getoor, L. PAC-Bayesian Collective Stability. In *Proc. of the 17th Int. Conf. on Artificial Intelligence and Statistics (AISTATS 14)*, 2014.

Massart, P. About the constants in Talagrand's concentration inequalities for empirical processes. *The Annals of Probability*, 28(2):863–884, 2000.

Maurer, A. Concentration inequalities for functions of independent variables. *Random Structures and Algorithms*, 29:121–138, 2006.

McDiarmid, C. On the method of bounded differences. *Survey in Combinatorics*, pp. 148–188, 1989.

McDiarmid, C. and Reed, B. Concentration for Self-bounding Functions and an Inequality of Talagrand. *Random Structures and Algorithms*, 29(4):549–557, 2006.

Mohri, M. and Rostamizadeh, A. Stability Bounds for Non-i.i.d. Processes. In *Adv. in Neural Information Processing Systems 20*, pp. 1025–1032, 2008.

Mohri, M. and Rostamizadeh, A. Rademacher Complexity Bounds for Non-I.I.D. Processes. In *Adv. in Neural Information Processing Systems 21*, pp. 1097–1104, 2009.

Ralaivola, L., Szafranski, M., and Stempfel, G. Chromatic PAC-Bayes Bounds for non-IID Data. In *AISTATS 09: JMLR Workshop and Conference Proceedings*, volume 5, pp. 416–423, 2009.

Ralaivola, L., Szafranski, M., and Stempfel, G. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. *JMLR, Journal of Machine Learning Research*, pp. 1–30, 2010.

Samson, P.-M. Concentration of measure inequalities for Markov chains and Φ-mixing processes. *Annals of Probability*, 28(1):416–461, 2000.

Steinwart, I. and Christmann, A. Fast learning from non-i.i.d. observations. In *Adv. in Neural Information Processing Systems 22*, pp. 1768–1776, 2010.

Usunier, N., Amini, M.-R., and Gallinari, P. Generalization Error Bounds for Classifiers Trained with Interdependent Data. In *Adv. in Neural Information Processing Systems 18*, pp. 1369–1376, 2006.

Yu, B. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94–116, 1994.