Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes

Liva Ralaivola, Marie Szafranski, Guillaume Stempfel

▶ **To cite this version:**

Liva Ralaivola, Marie Szafranski, Guillaume Stempfel. Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes. 2009. hal-00415162v2

HAL Id: hal-00415162
https://hal.science/hal-00415162v2

Preprint submitted on 3 Jun 2010

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.

Journal of Machine Learning Research () Submitted ; Published

Chromatic PAC-Bayes Bounds for Non-IID Data: Applications to Ranking and Stationary β-Mixing Processes

Liva Ralaivola LIVA.RALAIVOLA@LIF.UNIV-MRS.FR
Marie Szafranski MARIE.SZAFRANSKI@LIF.UNIV-MRS.FR
Guillaume Stempfel GUILLAUME.STEMPFEL@LIF.UNIV-MRS.FR
Laboratoire d'Informatique Fondamentale de Marseille
CNRS, Aix-Marseille Universités
39, rue F. Joliot Curie, 13013 Marseille, France

Editor:

Abstract

PAC-Bayes bounds are among the most accurate generalization bounds for classifiers learned from independently and identically distributed (IID) data, and it is particularly so for margin classifiers: there have been recent contributions showing how practical these bounds can be either to perform model selection (Ambroladze et al., 2007) or even to directly guide the learning of linear classifiers (Germain et al., 2009). However, there are many practical situations where the training data show some dependencies and where the traditional IID assumption does not hold. Stating generalization bounds for such frameworks is therefore of the utmost interest, both from theoretical and practical standpoints. In this work, we propose the first – to the best of our knowledge – PAC-Bayes generalization bounds for classifiers trained on data exhibiting interdependencies. The approach undertaken to establish our results is based on the decomposition of a so-called dependency graph that encodes the dependencies within the data, in sets of independent data, thanks to graph *fractional covers*. Our bounds are very general, since being able to find an upper bound on the fractional chromatic number of the dependency graph is sufficient to get new PAC-Bayes bounds for specific settings. We show how our results can be used to derive bounds for ranking statistics (such as AUC) and classifiers trained on data distributed according to a stationary β-mixing process. In the way, we show how our approach seemlessly allows us to deal with U-processes. As a side note, we also provide a PAC-Bayes generalization bound for classifiers learned on data from stationary φ-mixing distributions.

Keywords: PAC-Bayes bounds, non IID data, ranking, U-statistics, mixing processes.

1. Introduction

1.1 Background

Recently, there has been much progress in the field of generalization bounds for classifiers, the most noticeable of which are Rademacher-complexity-based bounds (Bartlett and Mendelson, 2002; Bartlett et al., 2005), stability-based bounds (Bousquet and Elisseeff, 2002) and PAC-Bayes bounds (McAllester, 1999). PAC-Bayes bounds, introduced by McAllester (1999), and refined in several occasions (Seeger, 2002a; Langford, 2005; Audibert and Bousquet, 2007), are some of the most appealing advances from the tightness and accuracy points of view (an excellent monograph on the PAC-Bayesian framework is that of

© Liva Ralaivola, Marie Szafranski and Guillaume Stempfel.

Catoni (2007)). Among others, striking results have been obtained concerning Pac-Bayes bounds for linear classifiers: Ambroladze et al. (2007) showed that Pac-Bayes bounds are a viable route to do actual model selection; Germain et al. (2009) recently proposed to learn linear classifiers by directly minimizing the linear Pac-Bayes bound with conclusive results, while Langford and Shawe-taylor (2002) showed that under some margin assumption, the Pac-Bayes framework allows one to tightly bound not only the risk of the stochastic Gibbs classsifier (see below) but also the risk of the Bayes classifier. The variety of (algorithmic, theoretical, practical) outcomes that can be expected from original contributions in the Pac-Bayesian setting explains and justifies the increasing interest it generates.

1.2 Contribution

To the best of our knowledge, Pac-Bayes bounds have essentially been derived for the setting where the training data are *independently and identically distributed* (IID). Yet, being able to learn from non-IID data while having strong theoretical guarantees on the generalization properties of the learned classifier is an actual problem in a number of real world applications such as, e.g., bipartite ranking (and more generally k-partite ranking) or classification from sequential data. Here, we propose the first Pac-Bayes bounds for classifiers trained on non-IID data; they constitute a generalization of the IID Pac-Bayes bound and they are generic enough to provide a principled way to establish generalization bounds for a number of non-IID settings. To establish these bounds, we make use of simple tools from probability theory, convexity properties of some functions, and we exploit the notion of *fractional covers* of graphs (Schreinerman and Ullman, 1997). One way to get a high level view of our contribution is the following: fractional covers allow us to cope with the dependencies within the set of random variables at hand by providing a strategy to make (large) subsets of independent random variables on which the usual IID Pac-Bayes bound is applied. Note that we essentially provide bounds for the case of *identically and non-independently* distributed data; the additional results that we give in the appendix generalizes to *non-identically and non-independently* distributed data.

1.3 Related Results

We would like to mention that the idea of dealing with sums of interdependent random variables by separating them into subsets of independent variables to establish concentration inequalities dates back to the work of Hoeffding (1948, 1963) on U-statistics. Explicity using the notion of (fractional) covers – or equivalently, colorings – of graphs to derive such concentration inequalities has been proposed by Pemmaraju (2001) and Janson (2004) and later extended by Usunier et al. (2006) to deal with functions that are different from the sum. Just as Usunier et al. (2006), who used their concentration inequality to provide generalization bounds based on the *fractional Rademacher complexity*, we take the approach of decomposing a set of dependent random variables into subsets of dependent random variables a step beyond establishing concentration inequality to provide what we call *chromatic* Pac-Bayes generalization bounds.

The genericity of our bounds is illustrated in several ways. It allows us to derive generalization bounds on the ranking performance of scoring/ranking functions using two different performance measures, among which the *Area under the* Roc *curve* (Auc) . These

bounds are directly related to the work of Agarwal et al. (2005), Agarwal and Niyogi (2009), Clémençon et al. (2008) and Freund et al. (2003). Even if our bounds are obtained as simple specific instances of our generic Pac-Bayes bounds, they exhibit interesting peculiarities. Compared with the bound of Agarwal et al. (2005) and Freund et al. (2003), our Auc bound depends in a less stronger way on the *skew* (i.e. the imbalance between positive and negative data) of the distribution; besides it does not rest on (rank-)shatter coefficients/VC dimension that may sometimes be hard to assess accurately; in addition, our bound directly applies to (kernel-based) linear classifiers. Agarwal and Niyogi (2009) base their analysis of ranking performances on algorithmic stability, and the qualitative comparison of their bounds and ours is not straightforward because stability arguments are somewhat different from the arguments used for Pac-Bayes bounds (and other uniform bounds). As already observed by Janson (2004), coloring provides a way to generalize large deviation results based on U-statistics; this observation carries over when generalization bounds are considered, which allows us to draw a connection between the results we obtain and that of Clémençon et al. (2008).

Another illustration of the genericity of our approach deals with mixing processes. In particular, we show how our chromatic bounds can be used to easily derive new generalization bounds for β-mixing processes. Rademacher complexity based bounds for such type of processes have recently been established by Mohri and Rostamizadeh (2009). To the best of our knowledge, it is the first time that such a bound is given in the Pac-Bayes framework. The striking feature is that it is done at a very low price: the independent block method proposed by Yu (1994) directly gives a dependency graph whose chromatic number is straightforward to compute. As we shall see, this suffices to instantiate our chromatic bounds, which, after simple calculations, leads to appropriate generalization bound. For sake of completeness, we also provide a Pac-Bayes bound for stationary φ-mixing processes; it is based on a different approach and its presentation is postponed to the appendix together with the tools that allows us to derive it.

1.4 Organization of the Paper

The paper is organized as follows. Section 2 recalls the standard IID Pac-Bayes bound. Section 3 introduces the notion of fractional covers and states the new chromatic Pac-Bayes bounds, which rely on the fractional chromatic number of the *dependency graph* of the data at hand. Section 4 provides specific versions of our bounds for the case of IID data, ranking and stationary β-mixing processes, giving rise to original generalization bounds. A Pac-Bayes bound for stationary φ-mixing based on arguments different from the chromatic Pac-Bayes bound is provided, in the appendix.

2. IID Pac-Bayes Bound

We introduce notation that will hold from here on. We mainly consider the problem of binary classification over the *input space* $\mathcal{X}$ and we denote the set of possible labels as $\mathcal{Y} = \{-1, +1\}$ (for the case of ranking described in section 4, we we use $\mathcal{Y} = \mathcal{R}$); $\mathcal{Z}$ denotes the product space $\mathcal{X} \times \mathcal{Y}$. $\mathcal{H} \subseteq \mathcal{R}^{\mathcal{X}}$ is a family of real valued classifiers defined on $\mathcal{X}$: for $h \in \mathcal{H}$, the predicted output of $x \in \mathcal{X}$ is given by $\mathrm{sign}(h(x))$, where $\mathrm{sign}(x) = +1$ if $x \geq 0$ and -1 otherwise. D is a probability distribution defined over $\mathcal{Z}$ and $\mathbf{D}_m$ denotes

the distribution of an m-sample; for instance, $\mathbf{D}_m = \otimes_{i=1}^m D = D^m$ is the distribution of an IID sample $\mathbf{Z} = \{Z_i\}_{i=1}^m$ of size m ($Z_i \sim D$, $i = 1 \dots m$). P and Q are distributions over $\mathcal{H}$. For any positive integer m, $[m]$ stands for $\{1, \dots, m\}$.

The IID Pac-Bayes bound, can be stated as follows (McAllester, 2003; Seeger, 2002a; Langford, 2005).

Theorem 1 (IID Pac-Bayes Bound) *$\forall D$, $\forall \mathcal{H}$, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m = D^m$, the following holds:*

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q(\mathbf{Z}) || e_Q) \leq \frac{1}{m}\left[\mathrm{KL}(Q||P) + \ln \frac{m+1}{\delta}\right]. \tag{1}$$

This theorem provides a generalization error bound for the *Gibbs classifier* g_Q: given a distribution Q, this stochastic classifier predicts a class for $\mathbf{x} \in \mathcal{X}$ by first drawing a hypothesis h according to Q and then outputting $\mathrm{sign}(h(\mathbf{x}))$. Here, $\hat{e}_Q$ is the empirical error of g_Q on an IID sample $\mathbf{Z}$ of size m and e_Q is its true error:

$$\begin{aligned} \hat{e}_Q(\mathbf{Z}) &:= \mathbb{E}_{h\sim Q} \frac{1}{m}\sum_{i=1}^m r(h, Z_i) = \mathbb{E}_{h\sim Q}\hat{R}(h, \mathbf{Z}) \quad &\text{with } \hat{R}(h, \mathbf{Z}) := \tfrac{1}{m}\textstyle\sum_{i=1}^m r(h, Z_i) \\ e_Q &:= \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \hat{e}_Q(\mathbf{Z}) = \mathbb{E}_{h\sim Q} R(h) &\text{with } R(h) := \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\hat{R}(h, \mathbf{Z}), \end{aligned} \tag{2}$$

where, for $Z = (X, Y)$,

$$r(h, Z) := \mathbb{I}_{Yh(X)<0}.$$

Note that we will use this binary 0-1 risk function r throughout the paper and that a generalization of our results to bounded real-valued risk functions is given in appendix. Since $\mathbf{Z}$ is an (independently) identically distributed sample, we have

$$R(h) = \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\hat{R}(h, \mathbf{Z}) = \mathbb{E}_{Z\sim D} r(h, Z). \tag{3}$$

For $p, q \in [0, 1]$, $\mathrm{kl}(q||p)$ is the Kullback-Leibler divergence between the Bernoulli distributions with probabilities of success q and p, and $\mathrm{KL}(Q||P)$ is the Kullback-Leibler divergence between Q and P:

$$\begin{aligned} \mathrm{kl}(q||p) &:= q \ln \frac{q}{p} + (1-q)\ln\frac{1-q}{1-p} \\ \mathrm{KL}(Q||P) &:= \mathbb{E}_{h\sim Q} \ln \frac{Q(h)}{P(h)}, \end{aligned}$$

where $\mathrm{kl}(0||0) = \mathrm{kl}(1||1) = 0$. All along, we assume that the posteriors are absolutely continuous with respect to their corresponding priors.

It is straightforward to see that the mapping $\mathrm{kl}_q : t \mapsto \mathrm{kl}(q||q+t)$ is strictly increasing for $t \in [0, 1-q)$ and therefore defines a bijection from $[0, 1-q)$ to $\mathcal{R}_+$: we denote by kl_q^{-1} its inverse. Then, as pointed out by Seeger (2002a), the function $\mathrm{kl}^{-1} : (q, \varepsilon) \mapsto \mathrm{kl}^{-1}(q, \varepsilon) = \mathrm{kl}_q^{-1}(\varepsilon)$ is well-defined over $[0, 1) \times \mathcal{R}^+$, and, by definition:

$$t \geq \mathrm{kl}^{-1}(q, \varepsilon) \Leftrightarrow \mathrm{kl}(q||q+t) \geq \varepsilon.$$

This makes it possible to rewrite bound (1) in a more 'usual' form:

$$\forall Q,\ e_Q \leq \hat{e}_Q(\mathbf{Z}) + \mathrm{kl}^{-1}\left(\hat{e}_Q(\mathbf{Z}), \frac{1}{m}\left[\mathrm{KL}(Q||P) + \ln\frac{m+1}{\delta}\right]\right). \tag{4}$$

We observe that even if bounds (1) and (4) apply to the risk e_Q of the stochastic classifier g_Q, a straightforward argument gives that, if b_Q is the (deterministic) Bayes classifier such that $b_Q(x) = \mathrm{sign}(\mathbb{E}_{h\sim Q}h(x))$, then $R(b_Q) = \mathbb{E}_{Z\sim D}r(b_Q, Z) \leq 2e_Q$ (see for instance (Herbrich and Graepel, 2001)). Langford and Shawe-taylor (2002) show that under some margin assumption, $R(b_Q)$ can be bounded even more tightly.

3. Chromatic Pac-Bayes Bounds

The problem we focus on is that of generalizing Theorem 1 to the situation where there may exist probabilistic dependencies between the elements Z_i of $\mathbf{Z} = \{Z_i\}_{i=1}^m$ while the marginal distributions of the Z_i's are identical. As announced before, we provide Pac-Bayes bounds for classifiers trained on identically but not independently distributed data. These results rely on properties of a dependency graph that is built according to the dependencies within $\mathbf{Z}$. Before stating our new bounds, we thus introduce the concepts of graph theory that will play a role in their statements.

3.1 Dependency Graph, Fractional Covers

Definition 2 (Dependency Graph) *Let $\mathbf{Z} = \{Z_i\}_{i=1}^m$ be a set of m random variables taking values in some space $\mathcal{Z}$. The* dependency graph $\Gamma(\mathbf{Z}) = (V, E)$ *of $\mathbf{Z}$ is such that:*

- *the set of vertices V of $\Gamma(\mathbf{Z})$ is $V = [m]$;*
- *$(i, j) \notin E$ (there is no edge between i and j) $\Leftrightarrow$ Z_i and Z_j are independent.*

Definition 3 (Fractional Covers, Schreinerman and Ullman (1997)) *Let $\Gamma = (V, E)$ be an undirected graph, with $V = [m]$.*

- *$C \subseteq V$ is* independent *if the vertices in C are independent (no two vertices in C are connected).*
- *$\mathbf{C} = \{C_j\}_{j=1}^n$, with $C_j \subseteq V$, is a* proper cover *of V if each C_j is independent and $\bigcup_{j=1}^n C_j = V$. It is* exact *if $\mathbf{C}$ is a partition of V. The size of $\mathbf{C}$ is n.*
- *$\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, with $C_j \subseteq V$ and $\omega_j \in [0, 1]$, is a* proper exact fractional cover *of V if each C_j is independent and $\forall i \in V$, $\sum_{j=1}^n \omega_j \mathbb{I}_{i\in C_j} = 1$; $\omega(\mathbf{C}) = \sum_{j=1}^n \omega_i$ is the* chromatic weight *of $\mathbf{C}$.*
- *The (fractional) chromatic number $\chi(\Gamma)$ ($\chi^*(\Gamma)$) is the minimum size (chromatic weight) over all proper exact (fractional) covers of Γ*

A cover is a fractional cover such that all the weights ω_i are equal to 1 (and all the results we state for fractional covers apply to the case of covers). If n is the size of a cover, it means

that the nodes of the graph at hand can be colored with n colors in a way such that no two adjacent nodes receive the same color.

The problem of computing the (fractional) chromatic number of a graph is NP-hard (Schreinerman and Ullman, 1997). However, for some particular graphs as those that come from the settings we study in Section 4, this number can be evaluated precisely. If it cannot be evaluated, it can be upper bounded using the following property.

Property 1 (Schreinerman and Ullman (1997)) *Let $\Gamma = (V, E)$ be a graph. Let $c(\Gamma)$ be the* clique number *of Γ, i.e. the order of the largest clique in Γ. Let $\Delta(\Gamma)$ be the maximum degree of a vertex in Γ. We have the following inequalities:*

$$1 \leq c(\Gamma) \leq \chi^*(\Gamma) \leq \chi(\Gamma) \leq \Delta(\Gamma) + 1.$$

In addition, $1 = c(\Gamma) = \chi^*(\Gamma) = \chi(\Gamma) = \Delta(\Gamma) + 1$ if and only if Γ is totally disconnected.

If $\mathbf{Z} = \{Z_i\}_{i=1}^m$ is a set of random variables over $\mathcal{Z}$ then a (fractional) proper cover of $\Gamma(\mathbf{Z})$, splits $\mathbf{Z}$ into subsets of independent random variables. This is a crucial feature to establish our results. In addition, we can see $\chi^*(\Gamma(\mathbf{Z}))$ and $\chi(\Gamma(\mathbf{Z}))$ as measures of the amount of dependencies within $\mathbf{Z}$.

The following lemma (Lemma 3.1 in (Janson, 2004)) will be very useful in the following.

Lemma 4 *If $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$ is an exact fractional cover of $\Gamma = (V, E)$, with $V = [m]$, then*

$$\forall \mathbf{t} \in \mathcal{R}^m, \sum_{i=1}^m t_i = \sum_{j=1}^n \omega_j \sum_{k \in C_j} t_k.$$

In particular, $m = \sum_{j=1}^n \omega_j |C_j|$.

3.2 Chromatic PAC-Bayes Bounds

We now provide new PAC-Bayes bounds for classifiers trained on samples $\mathbf{Z}$ drawn from distributions $\mathbf{D}_m$ where dependencies exist. We assume these dependencies are fully determined by $\mathbf{D}_m$ and we define the dependency graph $\Gamma(\mathbf{D}_m)$ of $\mathbf{D}_m$ to be $\Gamma(\mathbf{D}_m) = \Gamma(\mathbf{Z})$. As said before, the marginal distributions of $\mathbf{D}_m$ along each coordinate are the same and are equal to some distribution D.

We introduce additional notation. $\textsc{Pefc}(\mathbf{D}_m)$ is the set of proper exact fractional covers of $\Gamma(\mathbf{D}_m)$. Given a cover $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$, we use the following notation:

- $\mathbf{Z}^{(j)} = \{Z_k\}_{k \in C_j}$;
- $\mathbf{D}_m^{(j)}$, the distribution of $\mathbf{Z}^{(j)}$: it is equal to $D^{|C_j|} = \otimes_{i=1}^{|C_j|} D$ (C_j is independent);
- $\boldsymbol{\alpha} = (\alpha_j)_{1 \leq j \leq n}$ with $\alpha_j = \omega_j / \omega(\mathbf{C})$: we have $\alpha_j \geq 0$ and $\sum_j \alpha_j = 1$;
- $\boldsymbol{\pi} = (\pi_j)_{1 \leq j \leq n}$, with $\pi_j = \omega_j |C_j| / m$: we have $\pi_j \geq 0$ and $\sum_j \pi_j = 1$ (cf. Lemma 4).

In addition, $\mathbf{P}_n$ and $\mathbf{Q}_n$ denote distributions over $\mathcal{H}^n$, P_n^j and Q_n^j are the marginal distributions of $\mathbf{P}_n$ and $\mathbf{Q}_n$ with respect to the jth coordinate, respectively.

We can now state our main results.

Theorem 5 (Chromatic PAC-Bayes Bound (I)) $\forall \mathbf{D}_m$, $\forall \mathcal{H}$, $\forall \delta \in (0,1]$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \text{PEFC}(\mathbf{D}_m)$, $\forall \mathbf{P}_n$, *with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$, *the following holds:*

$$\forall \mathbf{Q}_n,\ \text{kl}(\bar{e}_{\mathbf{Q}_n}(\mathbf{Z})||e_{\mathbf{Q}_n}) \leq \frac{\omega}{m}\left[\sum_{j=1}^n \alpha_j \,\text{KL}(Q_n^j||P_n^j) + \ln\frac{m+\omega}{\delta\omega}\right], \tag{5}$$

where ω *stands for* $\omega(\mathbf{C})$, *and*

$$\bar{e}_{\mathbf{Q}_n}(\mathbf{Z}) := \sum_{j=1}^n \pi_j \mathbb{E}_{h\sim Q_n^j} \hat{R}(h, \mathbf{Z}^{(j)}),$$
$$e_{\mathbf{Q}_n} := \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}(\mathbf{Z}).$$

Proof Deferred to Section 3.4. ■

We would like to emphasize that the same type of result – using the same proof techniques – can be obtained if simple (i.e. not exact nor proper) fractional covers are considered. However, as we shall see, the 'best' (in terms of tightness) bound is achieved for covers from the set of proper exact fractional covers, and this is the reason why we have stated Theorem 5 with a restriction to this particular set of covers.

The empirical quantity $\bar{e}_{\mathbf{Q}_n}(\mathbf{Z})$ is a weighted average of the empirical errors on $\mathbf{Z}^{(j)}$ of Gibbs classifiers with respective distributions Q_n^j. The following proposition characterizes $e_{\mathbf{Q}_n} = \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}(\mathbf{Z})$.

Proposition 6 $\forall \mathbf{D}_m$, $\forall \mathcal{H}$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \text{PEFC}(\mathbf{D}_m)$, $\forall \mathbf{Q}_n$: $e_{\mathbf{Q}_n} = \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}(\mathbf{Z})$ *is the error of the Gibbs classifier based on the mixture of distributions* $Q^{\boldsymbol{\pi}} = \sum_{j=1}^n \pi_j Q_n^j$.

Proof From the definition of $\boldsymbol{\pi}$, $\pi_j \geq 0$ and $\sum_{j=1}^n \pi_j = 1$. Thus,

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \bar{e}_{\mathbf{Q}_n}(\mathbf{Z}) &= \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \sum_j \pi_j \mathbb{E}_{h\sim Q_n^j} \hat{R}(h, \mathbf{Z}^{(j)}) \\
&= \sum_j \pi_j \mathbb{E}_{h\sim Q_j} \mathbb{E}_{\mathbf{Z}^{(j)}\sim\mathbf{D}_m^{(j)}} \hat{R}(h, \mathbf{Z}^{(j)}) && \text{(marginalization)} \\
&= \sum_j \pi_j \mathbb{E}_{h\sim Q_n^j} R(h) && (\mathbb{E}_{\mathbf{Z}^{(j)}\sim\mathbf{D}_m^{(j)}} \hat{R}(h, \mathbf{Z}^{(j)}) = R(h), \forall j) \\
&= \mathbb{E}_{h\sim\pi_1 Q_n^1 + \ldots + \pi_j Q_n^j} R(h) = \mathbb{E}_{h\sim Q^{\boldsymbol{\pi}}} R(h).
\end{aligned}$$

Where, in the third line, we have used the fact that the variables in $\mathbf{Z}^{(j)}$ are identically distributed (by assumption, they are IID). ■

Remark 7 *The prior* $\mathbf{P}_n$ *and the posterior* $\mathbf{Q}_n$ *enter into play in Proposition 6 and Theorem 5 through their marginals only. This advocates for the following learning scheme. Given a cover and a (possibly factorized) prior* $\mathbf{P}_n$, *look for a factorized posterior* $\mathbf{Q}_n = \otimes_{j=1}^n Q_j$ *such that each* Q_j *independently minimizes the usual* IID PAC*-Bayes bound given in Theorem 1 on each* $\mathbf{Z}^{(j)}$. *Then make predictions according to the Gibbs classifier defined with respect to* $Q^{\boldsymbol{\pi}} = \sum_j \pi_j Q_j$.

The following theorem gives a result that readily applies without choosing a specific cover.

Theorem 8 (Chromatic PAC-Bayes Bound (II)) $\forall \mathbf{D}_m, \forall \mathcal{H}, \forall \delta \in (0,1], \forall P$, *with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$, *the following holds*

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q(\mathbf{Z})||e_Q) \leq \frac{\chi^*}{m}\left[\mathrm{KL}(Q||P) + \ln\frac{m+\chi^*}{\delta\chi^*}\right], \tag{6}$$

where χ^* *is the fractional chromatic number of* $\Gamma(\mathbf{D}_m)$, *and where* $\hat{e}_Q(\mathbf{Z})$ *and* e_Q *are as in (2).*

Proof This theorem is just a particular case of Theorem 5. Assume that $\mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$ such that $\omega(C) = \chi^*(\Gamma(\mathbf{D}_m))$, $\mathbf{P}_n = \otimes_{j=1}^n P = P^n$ and $\mathbf{Q}_n = \otimes_{j=1}^n Q = Q^n$, for some P and Q.

For the right-hand side of (6), it directly comes that

$$\sum_j \alpha_j \,\mathrm{KL}(Q_n^j||P_n^j) = \sum_j \alpha_j \,\mathrm{KL}(Q||P) = \mathrm{KL}(Q||P).$$

It then suffices to show that $\bar{e}_{\mathbf{Q}_n}(\mathbf{Z}) = \hat{e}_Q(\mathbf{Z})$:

$$\begin{aligned}
\bar{e}_{\mathbf{Q}_n}(\mathbf{Z}) &= \sum_j \pi_j \mathbb{E}_{h\sim Q_n^j} \hat{R}(h, \mathbf{Z}^{(j)}) = \sum_j \pi_j \mathbb{E}_{h\sim Q} \hat{R}(h, \mathbf{Z}^{(j)}) \\
&= \frac{1}{m}\sum_j \omega_j |C_j| \mathbb{E}_{h\sim Q} \frac{1}{|C_j|}\sum_k r(h, Z_k) && (\pi_j = \tfrac{\omega_j|C_j|}{m}, \forall j) \\
&= \mathbb{E}_{h\sim Q}\frac{1}{m}\sum_j \omega_j \sum_k r(h, Z_k) \\
&= \mathbb{E}_{h\sim Q}\frac{1}{m}\sum_i r(h, Z_i) && \text{(cf. Lemma 4)} \\
&= \mathbb{E}_{h\sim Q}\hat{R}(h, \mathbf{Z}) = \hat{e}_Q(\mathbf{Z}).
\end{aligned}$$

■

A few comments are in order.

- A χ^* worsening. This theorem says that even in the case of non IID data, a PAC-Bayes bound very similar to the IID PAC-Bayes bound (1) can be stated, with a worsening (since $\chi^* \geq 1$) proportional to χ^*, i.e proportional to the amount of dependencies in the data. In addition, the new PAC-Bayes bounds is valid with any priors and posteriors, without the need for these distributions to depend on the chosen cover (as is the case with the more general Theorem 5).

- χ^*: the optimal constant. Among all elements of $\textsc{Pefc}(\mathbf{D}_m)$, χ^* is the best constant achievable in terms of the tightness of the bound (6) on e_Q: getting an optimal coloring gives rise to an 'optimal' bound. Indeed, it suffices to observe that the right-hand side of (5) is decreasing with respect to ω when all Q_n^j are identical (we let the reader check that). As χ^* is the smallest chromatic weight, it gives the tightest bound.

(a) $\Gamma_{\text{1-edge}}$ (b) Γ_u

Figure 1: Γ_u is the subgraph induced by $\Gamma_{\text{1-edge}}$ – which contains only one edge, between u and v – when u is removed: it might be preferable to consider the distribution corresponding to Γ_u in Theorem 8 instead of the distribution defined wrt $\Gamma_{\text{1-edge}}$, since $\chi^*(\Gamma_{\text{1-edge}}) = 2$ and $\chi^*(\Gamma_u) = 1$ (see text for detailed comments).

- $\Gamma(\mathbf{D}_m)$ vs. induced subgraphs. If $\mathbf{s} \subseteq [m]$ and $\mathbf{Z_s} = \{Z_s : s \in \mathbf{s}\}$, it is obvious that Theorem 8 holds for $|\mathbf{s}|$-samples drawn from the marginal distribution $\mathbf{D_s}$ of $\mathbf{Z_s}$. Considering only $\mathbf{Z_s}$ amounts to working with the subgraph $\Gamma(\mathbf{D_s})$ of $\Gamma(\mathbf{D}_m)$ induced by the vertices in $\mathbf{s}$: this might provide a better bound in situations where $\chi^*(\mathbf{D_s})/|\mathbf{s}|$ is smaller than $\chi^*(\mathbf{D}_m)/m$ (this is not guaranteed, however, because the empirical error $\hat{e}_Q(\mathbf{Z_s})$ computed on $\mathbf{Z_s}$ might be larger than $\hat{e}_Q(\mathbf{Z})$). To see this, consider a graph $\Gamma_{\text{1-edge}} = (V, E)$ of m vertices where $|E| = 1$, i.e. there are only two nodes, say u and v, that are connected (see Figure 1). The fractional chromatic number $\chi^*_{\text{1-edge}}$ of $\Gamma_{\text{1-edge}}$ is 2 (u and v must use distinct colors) while the (fractional) chromatic number χ^*_u of the subgraph Γ_u of $\Gamma_{\text{1-edge}}$ obtained by removing u is 1: $\chi^*_{\text{1-edge}}$ is twice as big as χ^*_u while the number of nodes only differ by 1 and, for large m, this ratio roughly carries over for $\chi^*_{\text{1-edge}}/m$ and $\chi^*_u/(m-1)$.

This last comment outlines that considering a subset of $\mathbf{Z}$, or, equivalently, a subgraph of $\Gamma(\mathbf{D}_m)$, in (6), might provide a better generalization bound. However, it is assumed that the choice of the subgraph is done *before* computing the bound: the bound does only hold with probability $1-\delta$ for the chosen subgraph. To alleviate this and provide a bound that takes advantage of several induced subgraphs, we have the following proposition:

Proposition 9 *Let* $\{m\}^{\#k}$ *denote* $\{\mathbf{s} : \mathbf{s} \subseteq [m], |\mathbf{s}| = m-k\}$. $\forall \mathbf{D}_m$, $\forall \mathcal{H}$, $\forall k \in [m]$, $\forall \delta \in (0,1]$, $\forall P$, *with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$: $\forall Q$,

$$e_Q \leq \min_{\mathbf{s}\in\{m\}^{\#k}} \left\{ \hat{e}_Q(\mathbf{Z_s}) + \mathrm{kl}^{-1}\left(\hat{e}_Q(\mathbf{Z_s}), \frac{\chi^*_{\mathbf{s}}}{|\mathbf{s}|}\left[\mathrm{KL}(Q||P) + \ln\frac{|\mathbf{s}|+\chi^*_{\mathbf{s}}}{\chi^*_{\mathbf{s}}} + \ln\binom{m}{k} + \ln\frac{1}{\delta}\right]\right)\right\}. \quad (7)$$

where $\chi^*_{\mathbf{s}}$ *is the fractional chromatic number of* $\Gamma(\mathbf{D_s})$, *and where* $\hat{e}_Q(\mathbf{Z_s})$ *is the empirical error of the Gibbs classifier* g_Q *on* $\mathbf{Z_s}$, *that is:* $\hat{e}_Q(\mathbf{Z_s}) = \mathbb{E}_{h\sim Q}\hat{R}(h, \mathbf{Z_s})$.

Proof Simply apply the union bound to equation (6) of Theorem 8: for fixed k, there are $\binom{m}{m-k} = \binom{m}{k}$ subgraphs and using $\delta/\binom{m}{k}$ makes the bound hold with probability $1-\delta$ for all possible $\binom{m}{k}$ subgraphs (simultaneously). Making use of the form (4) gives the result. ■

This bound is particularly useful when, for some small k, there exists a subset $\mathbf{s} \subseteq \{m\}^{\#k}$ such that the induced subgraph $\Gamma(\mathbf{D_s})$, which has k fewer nodes than $\Gamma(\mathbf{D}_m)$, has a fractional chromatic number $\chi^*_{\mathbf{s}}$ that is smaller than $\chi^*(\mathbf{D}_m)$ (as is the case with the graph $\Gamma_{\text{1-edge}}$ of

Figure 1, where $k = 1$). Obtaining a similar result that holds for subgraphs associated with sets $\mathbf{s}$ of sizes *larger or equal* to $m - k$ is possible by replacing $\ln \binom{m}{k}$ with $\ln \sum_{\kappa=0}^{k} \binom{m}{\kappa}$ in the bound (in that case, k should be kept small enough with respect to m, e.g. $k = \mathcal{O}_m(1)$, to ensure that the resulting bound still goes down to zero when $m \to \infty$).

3.3 On the Relevance of Fractional Covers

One may wonder whether using the fractional cover framework is the only way to establish a result similar to the one provided by Theorem 5. Of course, this is not the case and one may imagine other ways of deriving closely related results without mentioning the idea of fractional/cover coloring. (For instance, one may manipulate subsets of independent variables, assign weights to these subsets without referring to fractional covers, and arrive at results that are comparable to ours.)

However, if we assume that singling out independent sets of variables is the cornerstone of dealing with interdependent random variables, we find it enlightning to cast our approach within the rich and well-studied fractional cover/coloring framework. On the one hand, our objective of deriving tight bounds amounts to finding a decomposition of the set of random variables at hand into *few and large* independent subsets and taking the graph theory point of view, this obviously corresponds to a problem of graph coloring. Explicitly using the fractional cover/coloring argument allows us to directly benefit from the wealth of related results, such as Property 1 or, for instance, approaches as to how compute a cover or approximate the fractional chromatic number (e.g., linear programming). On the other hand, from a technical point of view, making use of the fractional cover argument allows us to preserve the simple structure of the proof of the classical IID PAC-Bayes bound to derive Theorem 5.

To summarize, the richness of the results on graph (fractional) coloring provides us with elegant tools to deal with a natural representation of the dependencies that may occur within a set of random variables. In addition, and as showed in this article, it is possible to seamlessly take advantage of these tools in the PAC-Bayesian framework (and probably in other bound-related frameworks).

3.4 Proof of Theorem 5

A proof in three steps, following the lines of the proofs given by Seeger (2002a) and Langford (2005) for the IID Pac-Bayes bound, can be provided.

Lemma 10 $\forall \mathbf{D}_m$, $\forall \delta \in (0, 1]$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^{n}$, $\forall \mathbf{P}_n$ *distribution over* $\mathcal{H}^n$, *with probability at least* $1 - \delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m$, *the following holds (here,* ω *stands for* $\omega(\mathbf{C})$*)*

$$\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n} \sum_{j=1}^{n} \alpha_j e^{|C_j|\,\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \leq \frac{m+\omega}{\delta\omega}, \tag{8}$$

where $\mathbf{h} = (h_1, \ldots, h_n)$ *is a random vector of hypotheses.*

Proof We first observe the following:

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\sum_j \alpha_j e^{|C_j|\,\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} &= \sum_j \alpha_j \mathbb{E}_{\mathbf{Z}^{(j)}\sim\mathbf{D}_m^{(j)}} e^{|C_j|\,\mathrm{kl}(\hat{R}(h,\mathbf{Z}^{(j)})||R(h))} \\
&\leq \sum_j \alpha_j(|C_j|+1) \qquad \text{(Lemma 20, Appendix)} \\
&= \frac{1}{\omega}\sum_j \omega_j(|C_j|+1) \\
&= \frac{m+\omega}{\omega}, \qquad \text{(Lemma 4)}
\end{aligned}$$

where using Lemma 20 is made possible by the fact that $\mathbf{Z}^{(j)}$ is an IID sample. Therefore,

$$\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m}\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_{j=1}^{n}\alpha_j e^{|C_j|\,\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \leq \frac{m+\omega}{\omega}.$$

According to Markov's inequality (Theorem 22, Appendix),

$$\mathbb{P}_{\mathbf{Z}}\left(\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_j \alpha_j e^{|C_j|\,\mathrm{kl}(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j))} \geq \frac{m+\omega}{\omega\delta}\right) \leq \delta.$$

■

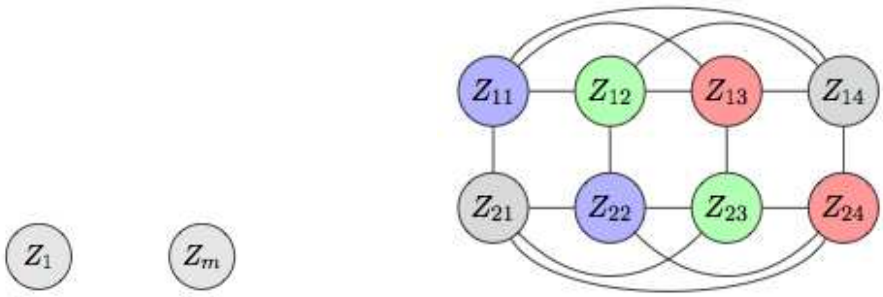
Lemma 11 *$\forall\mathbf{D}_m$, $\forall\mathbf{C}=\{(C_j,\omega_j)\}_{j=1}^n$, $\forall\mathbf{P}_n$, $\forall\mathbf{Q}_n$, with probability at least $1-\delta$ over the random draw of $\mathbf{Z}\sim\mathbf{D}_m$, the following holds*

$$\frac{m}{\omega}\sum\nolimits_{j=1}^{n}\pi_j\mathbb{E}_{h\sim Q_n^j}\,\mathrm{kl}(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)) \leq \sum\nolimits_{j=1}^{n}\alpha_j\,\mathrm{KL}(Q_n^j||P_n^j) + \ln\frac{m+\omega}{\delta\omega}. \tag{9}$$

Proof It suffices to use Jensen's inequality (Theorem 21, Appendix) with ln and the fact that $\mathbb{E}_{X\sim P}f(X)=\mathbb{E}_{X\sim Q}\frac{P(X)}{Q(X)}f(X)$, for all f,P,Q. Therefore, $\forall\mathbf{Q}_n$:

$$\begin{aligned}
\ln\mathbb{E}_{\mathbf{h}\sim\mathbf{P}_n}\sum_j \alpha_j e^{|C_j|\,\mathrm{kl}\left(\hat{R}(h_j,\mathbf{Z}^{(j)})||R(h_j)\right)} &= \ln\sum_j \alpha_j \mathbb{E}_{h\sim P_n^j} e^{|C_j|\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)} \\
&= \ln\sum_j \alpha_j \mathbb{E}_{h\sim Q_n^j}\frac{P_n^j(h)}{Q_n^j(h)} e^{|C_j|\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)} \\
&\geq \sum_j \alpha_j \mathbb{E}_{h\sim Q_n^j}\ln\left[\frac{P_n^j(h)}{Q_n^j(h)} e^{|C_j|\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right)}\right] \qquad \text{(Jensen's inequality)} \\
&= -\sum_j \alpha_j\,\mathrm{KL}(Q_n^j||P_n^j) + \sum_j \alpha_j|C_j|\mathbb{E}_{h\sim Q_n^j}\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right) \\
&= -\sum_j \alpha_j\,\mathrm{KL}(Q_n^j||P_n^j) + \frac{m}{\omega}\sum_j \pi_j \mathbb{E}_{h\sim Q_n^j}\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z}^{(j)})||R(h)\right).
\end{aligned}$$

Lemma 10 then gives the result. ■



(a) IID data (b) Bipartite ranking data

Figure 2: Dependency graphs for different settings described in section 4. Nodes of the same color are part of the same cover element; hence, they are probabilistically independent. (a) When the data are IID, the dependency graph is disconnected and the fractional number is $\chi^* = 1$; (b) a dependency graph obtained for bipartite ranking from a sample of 4 positive and 2 negative instances: $\chi^* = 4$.

Lemma 12 $\forall \mathbf{D}_m$, $\forall \mathbf{C} = \{(C_j, \omega_j)\}_{j=1}^n$, $\forall \mathbf{Q}_n$,, *the following holds*

$$\frac{m}{\omega} \sum_{j=1}^n \pi_j \mathbb{E}_{h \sim Q_n^j} \mathrm{kl}(\hat{R}(h, \mathbf{Z}^{(j)}) || R(h)) \geq \mathrm{kl}(\bar{e}_Q || e_Q).$$

Proof This simply comes from the convexity of $\mathrm{kl}(x, y)$ in (x, y) (Lemma 23, Appendix). This, in combination with Lemma 11, closes the proof of Theorem 5. ■

4. Applications

In this section, we provide instances of Theorem 8 for various settings; amazingly, they alllow us to easily derive Pac-Bayes generalization bounds for problems such as ranking and learning from stationary β-mixing processes. The theorems we provide here are all new Pac-Bayes bounds for different non-IID settings.

4.1 IID Case

The first case we are interested in is the IID setting. In this case, the training sample $\mathbf{Z} = \{(X_i, Y_i)\}_{i=1}^m$ is distributed according to $\mathbf{D}_m = D^m$ and the fractional chromatic number of $\Gamma(\mathbf{D}_m)$ is $\chi^* = 1$, since the dependency graph, depicted in Figure 2a is totally disconnected (see Property 1). Plugging in this value of χ^* in the bound of Theorem 8 gives the IID Pac-Bayes bound of Theorem 1. This emphasizes the fact that the standard Pac-Bayes bound is a special case of our more general results.

4.2 General Ranking and Connection to U-Statistics

Here, the learning problem of interest is the following. D is a distribution over $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} = \mathcal{R}$ and one looks for a ranking rule $h \in \mathcal{R}^{\mathcal{X} \times \mathcal{X}}$ that minimizes the *ranking risk* $R^{\mathrm{rank}}(h)$

defined as:

$$R^{\text{rank}}(h) := \mathbb{P}_{\substack{(X,Y)\sim D \\ (X',Y')\sim D}} ((Y - Y')h(X, X') < 0). \tag{10}$$

For a random pair (X, Y), Y can be thought of as a score that allows one to rank objects: given two pairs (X, Y) and (X', Y'), X has a higher rank (or is 'better') than X' if $Y > Y'$. The ranking rule h predicts X to be better than X' if $\text{sign}(h(X, X')) = 1$ and conversely. The objective of learning is to produce a rule h that makes as few misrankings as possible, as measured by (10). Given a finite IID (according to D) sample $\mathbf{S} = \{(X_i, Y_i)\}_{i=1}^{\ell}$ an unbiased estimate of $R^{\text{rank}}(h)$ is $\hat{R}^{\text{rank}}(h, \mathbf{S})$, with:

$$\hat{R}^{\text{rank}}(h, \mathbf{S}) := \frac{1}{\ell(\ell-1)} \sum_{i \neq j} \mathbb{I}_{(Y_i - Y_j)h(X_i, X_j) < 0} = \frac{1}{\ell(\ell-1)} \sum_{i \neq j} \mathbb{I}_{Y_{ij} h(X_i, X_j) < 0}, \tag{11}$$

where $Y_{ij} := (Y_i - Y_j)$. A natural question is to bound the ranking risk for any learning rule h given $\mathbf{S}$, where the difficulty is that (11) is a sum of identically but not independently random variables, namely the variables $\mathbb{I}_{Y_{ij} h(X_i, X_j)}$.

Let us define $X_{ij} := (X_i, X_j)$, $Z_{ij} := (X_{ij}, Y_{ij})$, and $\mathbf{Z} := \{Z_{ij}\}_{i \neq j}$. We note that the number ℓ of training data suffices to determine the structure of the dependency graph Γ_{rank} of $\mathbf{Z}$ and its distribution, which we denote $\mathbf{D}_{\ell(\ell-1)}$. Henceforth, we are clearly in the framework for the application of the chromatic Pac-Bayes bounds defined in the previous section. In particular, to instantiate Theorem 8 to the present ranking problem, we simply need to have at hand the value χ^*_{rank} – or an upper bound thereof – of the fractional chromatic number of Γ_{rank}. We claim that $\chi^*_{\text{rank}} \leq \ell(\ell-1)/\lfloor \ell/2 \rfloor$ where $\lfloor x \rfloor$ is the largest integer less than or equal to x. We provide the following new Pac-Bayes bound for the ranking risk:

Theorem 13 (Ranking Pac-Bayes bound) *$\forall D$ over $\mathcal{X} \times \mathcal{Y}$, $\forall \mathcal{H} \subseteq \mathcal{R}^{\mathcal{X} \times \mathcal{X}}$, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{S} \sim D^{\ell}$, the following holds*

$$\forall Q,\ \text{kl}(\hat{e}^{rank}_Q(\mathbf{S}) || e^{rank}_Q) \leq \frac{1}{\lfloor \ell/2 \rfloor} \left[\text{KL}(Q||P) + \ln \frac{\lfloor \ell/2 \rfloor + 1}{\delta} \right], \tag{12}$$

where

$$\hat{e}^{rank}_Q(\mathbf{S}) := \mathbb{E}_{h \sim Q} \hat{R}^{rank}(h, \mathbf{S})$$
$$e^{rank}_Q := \mathbb{E}_{\mathbf{S} \sim D^{\ell}} \hat{e}^{rank}_Q(\mathbf{S}).$$

Proof We essentially need to prove our claim on the bound on χ^*_{rank}. To do so, we consider a fractional cover of Γ_{rank} motivated by the theory of U-statistics (Hoeffding, 1948, 1963). $\hat{R}(h, \mathbf{S})$ is indeed a U-statistics of order 2 and it might be rewritten as a sum of IID blocks as follows

$$\hat{R}(h, \mathbf{S}) = \frac{1}{\ell(\ell-1)} \sum_{i \neq j} r(h, Z_{ij}) = \frac{1}{\ell!} \sum_{\sigma \in \Sigma_{\ell}} \frac{1}{\lfloor \ell/2 \rfloor} \sum_{i=1}^{\lfloor \ell/2 \rfloor} r\left(h, Z_{\sigma(i)\sigma(\lfloor \ell/2 \rfloor + i)}\right),$$

where Σ_{ℓ} is the set of permutations over $[\ell]$. The innermost sum is obviously a sum of IID random variables as no two summands share the same indices.

A proper exact fractional cover $\mathbf{C}_{\text{rank}}$ can be derived from this decomposition as[1]

$$\mathbf{C}_{\text{rank}} := \left\{ \left(C_\sigma := \left\{ Z_{\sigma(i)\sigma(\lfloor \ell/2 \rfloor + i)} \right\}_{i=1}^{\lfloor \ell/2 \rfloor}, \omega_\sigma := \frac{1}{(\ell-2)!\lfloor \ell/2 \rfloor} \right) \right\}_{\sigma \in \Sigma_\ell}.$$

Indeed, as remarked before, each C_σ is an independent set and each random variable Z_{pq} for $p \neq q$, appears in exactly $(\ell-2)! \times \lfloor \ell/2 \rfloor$ sets C_σ (for i fixed, the number of permutations σ such that $\sigma(i) = p$ and $\sigma(\lfloor \ell/2 \rfloor + i) = q$ is equal to $(\ell-2)!$, i.e. the number of permutations on $\ell - 2$ elements; as i can take $\lfloor \ell/2 \rfloor$ values, this gives the result). Therefore, $\forall p, q, p \neq q$:

$$\sum_{\sigma \in \Sigma_\ell} \omega_\sigma \mathbb{I}_{Z_{pq} \in C_\sigma} = \frac{1}{(\ell-2)!\lfloor \ell/2 \rfloor} \sum_{\sigma \in \Sigma_\ell} \mathbb{I}_{Z_{pq} \in C_\sigma} = \frac{1}{(\ell-2)!\lfloor \ell/2 \rfloor} \times (\ell-2)!\lfloor \ell/2 \rfloor = 1,$$

which proves that $\mathbf{C}_{\text{rank}}$ is a proper exact fractional cover. Its weight $\omega(\mathbf{C}_{\text{rank}})$ is

$$\omega(\mathbf{C}_{\text{rank}}) = \ell! \times \omega_\sigma = \frac{\ell(\ell-1))}{\lfloor \ell/2 \rfloor}.$$

Hence, from the definition of χ^*_{rank},

$$\chi^*_{\text{rank}} \leq \frac{\ell(\ell-1))}{\lfloor \ell/2 \rfloor}.$$

The theorem follows by an instantiation of Theorem 8 with $m := \ell(\ell-1)$ and the bound on χ^*_{rank} we have just proven. ■

To our knowledge, this is the first Pac-Bayes bound on the ranking risk, while a Rademacher-complexity based analysis was given by Clémençon et al. (2008). In the proof, we have used arguments from the analysis of U-processes, which allow us to easily derive a convenient fractional cover of the dependency graph of $\mathbf{Z}$. Note however that our framework still applies even if not all the Z_{ij}'s are known, as required if an analysis based on U-processes is undertaken. This is particularly handy in practical situations where one may only be given the values Y_{ij} – but *not* the values of Y_i and Y_j – for a limited number of (i, j) pairs (and not all the pairs).

An interesting question is to know how the so-called Hoeffding decomposition used by Clémençon et al. (2008) to establish fast rates of convergence for empirical ranking risk minimizers could be used to draw possibly tighter Pac-Bayes bounds. This would imply being able to appropriately take advantage of moments of order 2 in Pac-Bayes bounds, and a possible direction for that has been proposed by Lacasse et al. (2006). This is left for future work as it is not central to the present paper.

Of course, the ranking rule may be based on a scoring function $f \in \mathcal{R}^{\mathcal{X}}$ such that $h(X, X') = f(X) - f(X')$, in which case all the results that we state in terms of h can be stated similarly in terms of f. This is important to note from a practical point of view as it is probably more usual to learn functions defined over $\mathcal{X}$ rather than $\mathcal{X} \times \mathcal{X}$ (as is h).

Finally, we would like to stress that the bound on χ^*_{rank} that we have exhibited is actually rather tight. Indeed, it is straightforward to see that the clique number of Γ_{rank} is $2(\ell-1)$

1. Note that the cover defined here considers elements C_σ containing random variables themselves instead of their indices. This abuse of notation is made for sake of readability.

(the cliques are made of variables $\{Z_{ip}\}_p \bigcup \{Z_{pi}\}_p$ for every i), and according to Property 1, $2(\ell-1)$ is therefore a lower bound on χ^*_{rank}. If ℓ is even, then our bound on χ^*_{rank} is equal to $2(\ell-1)$ and so is χ^*_{rank}; if ℓ is odd, then our bound is 2ℓ.

4.3 Bipartite Ranking and a Bound on the Auc

A particular ranking setting is that of bipartite ranking, where $\mathcal{Y} = \{-1, +1\}$. Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ and D_{+1} (D_{-1}) be the class conditional distribution $D_{X|Y=+1}$ ($D_{X|Y=-1}$) with respect to D. In this setting (see, e.g. Agarwal et al. (2005)), one may be interested in controlling what we call the *bipartite misranking risk* $R^{\text{Auc}}(h)$ (the reason for the Auc superscript will become clear in the sequel), of a ranking rule $h \in \mathcal{R}^{\mathcal{X}\times\mathcal{X}}$ by

$$R^{\text{Auc}}(h) := \mathbb{P}_{\substack{X\sim D_{+1} \\ X'\sim D_{-1}}} (h(X, X') < 0). \tag{13}$$

Note that the relation between R^{Auc} and R^{rank} (cf. Equation (10)) can be made clear whenever the hypotheses h under consideration are such that $h(x, x')$ and $h(x', x)$ have opposite signs. In this situation, it is straightforward to see that

$$R^{\text{rank}}(h) = 2\mathbb{P}(Y = +1)\mathbb{P}(Y = -1)R^{\text{Auc}}(h).$$

Let $\mathbf{S} = \{(X_i, Y_i)\}_{i=1}^{\ell}$ be an IID sample distributed according to $\mathbf{D}_\ell = D^\ell$. The empirical bipartite ranking risk $\hat{R}^{\text{Auc}}(h, \mathbf{S})$ of h on $\mathbf{S}$ defined as

$$\hat{R}^{\text{Auc}}(h, \mathbf{S}) := \frac{1}{\ell^+\ell^-} \sum_{\substack{i:Y_i=+1 \\ j:Y_j=-1}} \mathbb{I}_{h(X_i,X_j)<0} \tag{14}$$

where ℓ^+ (ℓ^-) is the number of positive (negative) data in $\mathbf{S}$, estimates the fraction of pairs (X_i, X_j) that are incorrectly ranked incorrectly (given that $Y_i = +1$ and $Y_j = -1$) by h: it is an unbiased estimator of $R^{\text{Auc}}(h)$.

As before, h may be expressed in terms of a scoring function $f \in \mathcal{R}^{\mathcal{X}}$ such that $h(X, X') = f(X) - f(X')$, in which case (overloading notation):

$$R^{\text{Auc}}(f) = \mathbb{P}_{\substack{X\sim D_{+1} \\ X'\sim D_{-1}}} (f(X) < f(X')) \text{ and } \hat{R}^{\text{Auc}}(f, \mathbf{S}) = \frac{1}{\ell^+\ell^-} \sum_{\substack{i:Y_i=+1 \\ j:Y_j=-1}} \mathbb{I}_{f(X_i)<f(X_j)},$$

where we recognize in $\hat{R}^{\text{Auc}}(f, \mathbf{S})$ one minus the Area under the Roc curve, or Auc, of f on $\mathbf{S}$ (Agarwal et al., 2005; Cortes and Mohri, 2004), hence the Auc superscript in the name of the risk. As a consequence, providing a Pac-Bayes bound on $R^{\text{Auc}}(h)$ (or $R^{\text{Auc}}(f)$) amounts to providing a generalization (lower) bound on the Auc, which is a widely used measure in practice to evaluate the performance of a scoring function.

Let us define $X_{ij} := (X_i, X_j)$, $Z_{ij} := (X_{ij}, 1)$ and $\mathbf{Z} := \{Z_{ij}\}_{ij:Y_i=+1,Y_j=-1}$, i.e. $\mathbf{Z}$ is a sequence of pairs X_{ij} made of one positive example and one negative example. We then are once again in the framework defined earlier[2], i.e., the Z_{ij}'s share the same distribution but

2. The slight difference with what has been described above is that the dependency graph is now a random variable: it depends on the Y_i's. It is shown in the proof of Theorem 14 how this can be dealt with.

are dependent on each other, since Z_{ij} depends on $\{Z_{pq} : p = i \text{ or } q = j\}$ (see Figure 2). Note that in order to ease the reading of the present subsection, we make the implicit decomposition of training set $\mathbf{S}$ into $\mathbf{S} = \mathbf{S}^+ \cup \mathbf{S}^-$, where $\mathbf{S}^+$ (resp. $\mathbf{S}^-$) is made of the ℓ^+ (ℓ^-) positive (negative) data of $\mathbf{S}$; the size ℓ of $\mathbf{S}$ is therefore $\ell = \ell^+ + \ell^-$. This decomposition entails a separate reindexing of the positive (negative) data from 1 to ℓ^+ (from 1 to ℓ^-).

Building on Theorem 8, we have the following result:

Theorem 14 (Auc Pac-Bayes bound) *$\forall D$ over $\mathcal{X} \times \mathcal{Y}$, $\forall \mathcal{H} \subseteq \mathcal{R}^{\mathcal{X}\times\mathcal{X}}$, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{S} \sim D^\ell$, the following holds*

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q^{\mathrm{Auc}}(\mathbf{S})||e_Q^{\mathrm{Auc}}) \leq \frac{1}{\ell_{\min}} \left[\mathrm{KL}(Q||P) + \ln \frac{\ell_{\min} + 1}{\delta} \right], \tag{15}$$

where $\ell_{\min} = \min(\ell^+, \ell^-)$, and

$$\hat{e}_Q^{\mathrm{Auc}}(\mathbf{S}) := \mathbb{E}_{h\sim Q} \hat{R}^{\mathrm{Auc}}(h, \mathbf{S})$$
$$e_Q^{\mathrm{Auc}} := \mathbb{E}_{\mathbf{S}\sim D^\ell} \hat{e}_Q^{\mathrm{Auc}}(\mathbf{S}).$$

Proof The proof works in three steps and borrows ideas from Agarwal et al. (2005). The first two parts are necessary to deal with the fact that the dependency graph of $\mathbf{Z}$, as it depends on the random sample $\mathbf{S}$, does not have a deterministic structure.

Conditioning on $\mathbf{Y} = \mathbf{y}$. Let $\mathbf{y} \in \{-1, +1\}^\ell$ be a fixed vector and let $\ell_{\mathbf{y}}^+$ and $\ell_{\mathbf{y}}^-$ be the number of positive and negative labels in $\mathbf{y}$, respectively. We define the distribution $\mathbf{D}_{\mathbf{y}}$ as $\mathbf{D}_{\mathbf{y}} := \otimes_{i=1}^{\ell} D_{y_i}$; this is a distribution on $\mathcal{X}^\ell$. With a slight abuse of notation, $\mathbf{D}_{\mathbf{y}}$ will also be used to denote the distribution over $(\mathcal{X} \times \mathcal{Y})^\ell$ of samples $\mathbf{S} = \{(X_i, y_i)\}_{i=1}^{\ell}$ such that the sequence $\{X_i\}_{i=1}^{\ell}$ is distributed according to $\mathbf{D}_{\mathbf{y}}$. It is easy to check that $\forall h \in \mathcal{H}$, $\mathbb{E}_{\mathbf{S}\sim\mathbf{D}_{\mathbf{y}}} \hat{R}^{\mathrm{rank}}(h, \mathbf{S}) = R^{\mathrm{rank}}(h)$ (cf. equations (13) and (14)).

Given $\mathbf{S}$, if we define, as said earlier, $X_{ij} := (X_i, X_j)$, $Y_{ij} := 1$ and $Z_{ij} := (X_{ij}, Y_{ij})$, then $\mathbf{Z} := \{Z_{ij}\}_{i:y_i=1, j:y_j=-1}$ is a sample of identically distributed variables, each with distribution $D_{\pm 1} = D_{+1} \otimes D_{-1} \otimes \mathbf{1}$ over $\mathcal{X} \times \mathcal{X} \times \mathcal{Y}$, where $\mathcal{Y} = \{-1, +1\}$ and where $\mathbf{1}$ is the distribution that produces 1 with probability 1.

Letting $m = \ell_{\mathbf{y}}^+ \ell_{\mathbf{y}}^-$ we denote by $\mathbf{D}_{\mathbf{y},m}$ the distribution of the training sample $\mathbf{Z}$, within which interdependencies exist, as illustrated in Figure 2. Theorem 8 can thus be directly applied to classifiers trained on $\mathbf{Z}$, the structure of $\Gamma(\mathbf{D}_{\mathbf{y},m})$ and its corresponding fractional chromatic number $\chi_{\mathbf{y}}^*$ being completely determined by $\mathbf{y}$. Hence, letting $\mathcal{H} \subseteq \mathcal{R}^{\mathcal{X}\times\mathcal{X}}$, we have: $\forall \delta \in (0, 1]$, $\forall P$ over $\mathcal{H}$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_{\mathbf{y},m}$,

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q(\mathbf{Z})||e_Q) \leq \frac{\chi_{\mathbf{y}}^*}{m} \left[\mathrm{KL}(Q||P) + \ln \frac{m + \chi_{\mathbf{y}}^*}{\delta \chi_{\mathbf{y}}^*} \right],$$

where $\hat{e}_Q(\mathbf{Z}) = \mathbb{E}_{h\sim Q} \hat{R}(h, \mathbf{Z}) = \mathbb{E}_{h\sim Q} \sum_{ij} \mathbb{I}_{Y_{ij} h(Z_{ij}) < 0} = \mathbb{E}_{h\sim Q} \sum_{ij} \mathbb{I}_{h(Z_{ij}) < 0}$, which is exactly equal to $\hat{e}_Q^{\mathrm{Auc}}(\mathbf{S})$ (cf. (14)); likewise, $e_Q = \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_{\mathbf{y},m}} \hat{e}_Q(\mathbf{Z}) = \mathbb{E}_{\mathbf{S}\sim\mathbf{D}_{\mathbf{y}}} \hat{e}_Q^{\mathrm{Auc}}(\mathbf{S}) = e_Q^{\mathrm{Auc}}$. Hence, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{S} \sim \mathbf{D}_{\mathbf{y}}$,

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q^{\mathrm{Auc}}(\mathbf{S})||e_Q^{\mathrm{Auc}}) \leq \frac{\chi_{\mathbf{y}}^*}{m} \left[\mathrm{KL}(Q||P) + \ln \frac{m + \chi_{\mathbf{y}}^*}{\delta \chi_{\mathbf{y}}^*} \right]. \tag{16}$$

Unconditioning on Y. As proposed by Agarwal et al. (2005), let us call $\Phi(P, \mathbf{S}, \delta)$ the event (16); we just stated that $\forall \mathbf{y} \in \{-1, +1\}^\ell$, $\forall P$, $\forall \delta \in (0, 1]$, $\mathbb{P}_{\mathbf{S}\sim\mathbf{D}_\mathbf{y}}(\Phi(P, \mathbf{S}, \delta)) \geq 1 - \delta$, or, equivalently

$$\mathbb{P}_{\mathbf{S}\sim\mathbf{D}_\ell}(\neg\Phi(P, \mathbf{S}, \delta)|Y = \mathbf{y}) = \mathbb{P}_{\mathbf{S}\sim\mathbf{D}_\mathbf{y}}(\neg\Phi(P, \mathbf{S}, \delta)) < \delta,$$

i.e., the conditional (to $Y = \mathbf{y}$) probability of the event $\neg\Phi(P, \mathbf{S}, \delta)$ is bounded by δ. This directly implies that the unconditional probability of $\neg\Phi(P, \mathbf{S}, \delta)$ is bounded by δ as well:

$$\mathbb{P}_{\mathbf{S}\sim\mathbf{D}_\ell}(\neg\Phi(P, \mathbf{S}, \delta)) \leq \mathbb{P}_{\mathbf{S}\sim\mathbf{D}_\ell}(\neg\Phi(P, \mathbf{S}, \delta)|Y = \mathbf{y}) < \delta.$$

Hence, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{S} \sim \mathbf{D}_\ell$,

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q^{\mathrm{AUC}}||e_Q^{\mathrm{AUC}}) \leq \frac{\chi^*_\mathbf{S}}{m_\mathbf{S}}\left[\mathrm{KL}(Q||P) + \ln\frac{m_\mathbf{S} + \chi^*_\mathbf{S}}{\delta\chi^*_\mathbf{S}}\right]. \tag{17}$$

where $\chi^*_\mathbf{S}$ is the fractional chromatic number of the graph $\Gamma(\mathbf{Z})$, with $\mathbf{Z}$ defined from $\mathbf{S}$ as in the first part of the proof, where the observed (random) labels are now taken into account; here $m_\mathbf{S} = \ell^+\ell^-$, where ℓ^+ (ℓ^-) is the number of positive (negative) data in $\mathbf{S}$.

Computing the Fractional Chromatic Number. In order to finish the proof, it suffices to observe that, for $\mathbf{Z} = \{Z_{ij}\}_{ij}$, if $\ell_{\max} = \max(\ell^+, \ell^-)$, then the fractional chromatic number of $\Gamma(\mathbf{Z})$ is $\chi^* = \ell_{\max}$.

Indeed, the clique number of $\Gamma(\mathbf{Z})$ is $\ell_{\max}$ as for all $i = 1, \ldots, \ell^+$ ($j = 1, \ldots, \ell^-$), $\{Z_{ij} : j = 1, \ldots, \ell^-\}$ ($\{Z_{ij} : i = 1, \ldots, \ell^+\}$) defines a clique of order ℓ^- (ℓ^+) in $\Gamma(\mathbf{Z})$. Thus, from Property 1: $\chi \geq \chi^* \geq \ell_{\max}$.

A proper exact cover $\mathbf{C} = \{C_k\}_{k=1}^{\ell_{\max}}$ of $\Gamma(\mathbf{Z})$ can be constructed as follows. Suppose that $\ell_{\max} = \ell^+$, then $C_k = \{Z_{i\sigma_k(i)} : i = 1, \ldots, \ell^-\}$, with

$$\sigma_k(i) = (i + k - 2 \mod \ell^+) + 1,$$

is an independent set: no two variables Z_{ij} and Z_{pq} in C_k are such that $i = p$ or $j = q$. In addition, it is straightforward to check that $\mathbf{C}$ is indeed a cover of $\Gamma(\mathbf{Z})$. This cover is of size $\ell^+ = \ell_{\max}$, which means that it achieves the minimal possible weight over proper exact (fractional) covers since $\chi^* \geq \ell_{\max}$. Hence, $\chi^* = \chi = \ell_{\max}(= c(\Gamma))$. Plugging in this value of χ^* in (17), and noting that $m_\mathbf{S} = \ell_{\max}\ell_{\min}$ with $\ell_{\min} = \min(\ell^+, \ell^-)$, closes the proof. ■

We observe that in the theorem, the dependence on the skew of the sample is expressed in terms of $1/\min(\ell^+, \ell^-)$, whereas in the the works of Agarwal et al. (2005) and Usunier et al. (2005), the bound depends on the larger $1/\ell^+ + 1/\ell^-$.

The Pac-Bayes bound of Theorem 14 can be specialized to the case where $h(x, x') = f(x) - f(x')$ with $f \in \{x \mapsto w \cdot x : w \in \mathcal{X}\}$: f is therefore a linear scoring function and $h(x, x') = w \cdot (x - x')$. The ranking rule h is thus a linear classifier acting on the difference of its arguments (the next result we present therefore carries over to kernel classifiers). As proposed by Langford (2005), we may assume an isotropic Gaussian prior $P = \mathcal{N}(0, I)$ and a family of posteriors $Q_{w,\mu}$ parameterized by $w \in \overline{\mathcal{X}}$ and $\mu > 0$ such that $Q_{w,\mu}$ is $\mathcal{N}(\mu, 1)$ in the direction w and $\mathcal{N}(0, 1)$ in all perpendicular directions, we arrive at the following theorem:

Theorem 15 (Auc Linear Pac-Bayes bound) *$\forall \ell, \forall D$ over $\mathcal{X} \times \mathcal{Y}$, $\forall \delta \in (0,1]$, the following holds with probability at least $1-\delta$ over the draw of $\mathbf{S} \sim D^\ell$:*

$$\forall w, \mu > 0,\ \mathrm{kl}(\hat{e}^{\mathrm{Auc}}_{Q_{w,\mu}}(\mathbf{S}) || e^{\mathrm{Auc}}_{Q_{w,\mu}}) \leq \frac{1}{\ell_{\min}} \left[\frac{\mu^2}{2} + \ln \frac{\ell_{\min}+1}{\delta} \right].$$

Proof Straightforward from the bound of Langford (2005) and Theorem 14. ■

Note that this specific parametrization of Q could have been done in Theorem 13 as well. We arbitrarily choose to provide it for this Auc based bound as learning linear ranking rule by Auc minimization is a common approach (Ataman et al., 2006; Brefeld and Scheffer, 2005; Rakotomamonjy, 2004), and the presented result may be of practical interest (for model selection purpose, for instance) for a larger audience.

The bounds given in Theorem 14 and Theorem 15 are very similar to what we would get if applying IID Pac-Bayes bound to one (independent) element C_j of a minimal cover (i.e. its weight equals the fractional chromatic number) $\mathbf{C} = \{C_j\}_{j=1}^n$ such as the one we used in the proof of Theorem 14. This would imply the empirical error $\hat{e}^{\mathrm{rank}}_{\overline{Q}}$ to be computed on only one specific C_j and not all the C_j's simultaneously, as is the case for the new results. It turns out that, for proper exact fractional covers $\mathbf{C} = \{(C_j, \omega)\}_{j=1}^n$ with elements C_j having the same size, it is better, in terms of absolute moments of the empirical error, to assess it on the whole dataset, rather than on only one C_j. The following proposition formalizes this.

Proposition 16 *$\forall \mathbf{D}_m$, $\forall \mathcal{H}$, $\forall \mathbf{C} = \{(C_j, \omega_j\}_{j=1}^n \in \textsc{Pefc}(\mathbf{D}_m)$, $\forall Q$, $\forall r \in \mathcal{N}, r \geq 1$, if $|C_1| = \ldots = |C_n|$ then*

$$\mathbb{E}_{\mathbf{Z} \sim \mathbf{D}_m} |\hat{e}_Q(\mathbf{Z}) - e_Q|^r \leq \mathbb{E}_{\mathbf{Z}^{(j)} \sim \mathbf{D}_m^{(j)}} |\hat{e}_Q(\mathbf{Z}^{(j)}) - e_Q|^r, \forall j \in \{1, \ldots n\}.$$

Proof Using the convexity of $|\cdot|^r$ for $r \geq 1$, the linearity of $\mathbb{E}$ and the notation of section 3, for $\mathbf{Z} \sim \mathbf{D}_m$:

$$\begin{aligned} |\hat{e}_Q(\mathbf{Z}) - e_Q|^r &= |\sum_j \pi_j (\mathbb{E}_{h \sim Q} \hat{R}(h, \mathbf{Z}^{(j)}) - R(h))|^r \\ &\leq \sum_j \pi_j |\mathbb{E}_{h \sim Q} (\hat{R}(h, \mathbf{Z}^{(j)}) - R(h))|^r \\ &= \sum_j \pi_j |\hat{e}_Q(\mathbf{Z}^{(j)}) - e_Q|^r. \end{aligned}$$

Taking the expectation of both sides with respect to $\mathbf{Z}$ and noting that the random variables $|\hat{e}_Q(\mathbf{Z}^{(j)}) - e_Q|^r$, have the same distribution, gives the result. ■

This proposition upholds the idea of Pemmaraju (2001) to base the decomposition of a dependency graph on equitable coloring.

4.4 β-mixing Processes

Here, we provide a Pac-Bayes theorem for classifiers trained on data from a stationary β-mixing process, of which we recall some definitions, as formulated by Yu (1994) (see also, e.g., also Mohri and Rostamizadeh (2009)).

Definition 17 (Stationarity) *A sequence of random variables* $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *is* stationary *if, for any* t *and nonnegative integer* m *and* k*, the random subsequences* $(Z_t, \ldots, Z_{t+m})$ *and* $(Z_{t+k}, \ldots, Z_{t+m+k})$ *are identically distributed.*

Definition 18 (β-mixing process) *Let* $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ *be a stationary sequence of random variables. For any* $i, j \in \mathcal{Z} \cup \{-\infty, +\infty\}$*, let* σ_i^j *denote the* σ*-algebra generated by the random variables* Z_k*,* $i \leq k \leq j$*. Then, for any positive integer* k*, the* β*-mixing coefficient* $\beta(k)$ *of the stochastic process* $\mathbf{Z}$ *is defined as*

$$\beta(k) = \sup_{n \geq 1} \mathbb{E} \sup \left\{ |\mathbb{P}\left(A|\sigma_1^n\right) - \mathbb{P}\left(A\right)| : A \in \sigma_{n+k}^{+\infty} \right\}. \tag{18}$$

$\mathbf{Z}$ *is said to be* β*-mixing if* $\beta(k) \to 0$ *when* $k \to \infty$*.*

(Note there is an equivalent definition of the β-mixing coefficient based on finite partitions; see Yu (1994) for details.) Stationary β-mixing processes model a situation where the interdependence between the random variables at hand is temporal. When the process is mixing, it means that the strengh of dependence between variables weakens over times.

The bound that we propose is in the same vein as the one proposed by Mohri and Rostamizadeh (2009), with the difference that our bound is a Pac-Bayes bound and theirs a Rademacher-complexity-based bounds. In addition to being a new type of data-dependent bound for the case of stationary β-mixing process, we may anticipate that, in practical situations, our bound inherits the tightness of the IID Pac-Bayes bound (whereas, to the best of our knowledge, there is no evidence of such practicality for Rademacher-complexity-based bounds).

Let us state our generalization bound for classifiers trained on samples $\mathbf{Z}$ drawn from stationary β-mixing distributions.

Theorem 19 (β-mixing process Pac-Bayes bound) *Let* m *be a positive integer. Let* $\mathbf{D}^\beta$ *be a stationary* β*-mixing distribution over* $\mathcal{Z}$ *and* $\mathbf{D}_m^\beta$ *be the distribution of* m*-samples according to* $\mathbf{D}^\beta$*.* $\forall \mathcal{H} \subseteq \mathcal{R}^\mathcal{X}$*,* $\forall \mu, a \in \mathcal{N}$ *such that* $2\mu a = m$*,* $\forall \delta \in (2(\mu-1)\beta(a), 1]$*,* $\forall P$*, with probability at least* $1-\delta$ *over the random draw of* $\mathbf{Z} \sim \mathbf{D}_m^\beta$*, the following holds*

$$\forall Q,\ \mathrm{kl}(\hat{e}_Q^\beta(\mathbf{Z}) || e_Q^\beta) \leq \frac{1}{\mu}\left[\mathrm{KL}(Q||P) + \ln \frac{2(\mu+1)}{\delta - 2(\mu-1)\beta(a)}\right], \tag{19}$$

where

$$\hat{e}_Q^\beta(\mathbf{Z}) := \mathbb{E}_{h\sim Q}\hat{R}(h, \mathbf{Z}) = \mathbb{E}_{h\sim Q}\sum_{t=1}^{m} \mathbb{I}_{Y_t h(X_t)<0}$$
$$e_Q^\beta := \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}\hat{e}_Q^\beta(\mathbf{Z}).$$

Proof The proof makes use of the independent block decomposition proposed by Yu (1994), our chromatic Pac-Bayes bound of Theorem 8, and Corollary 24 (Appendix).

The chromatic bound for independent blocks. Let $\mathbf{Z} = \{Z_1, \ldots, Z_m\}$ be the random variables we have to deal with. If μ and a are two integers such that $2\mu a = m$ (we assume that m is even, if it is odd one may drop the last variable Z_m and work on a sample of size $m-1$). Then $\mathbf{Z}$ can be decomposed into two subsequences $\mathbf{Z}_0$ and $\mathbf{Z}_1$ as follows:

$$\mathbf{Z}_0 := \{\mathbf{Z}_0^s := (Z_{a(2s-2)+1}, \ldots, Z_{a(2s-2)+a}) : s \in [\mu]\},$$
$$\mathbf{Z}_1 := \{\mathbf{Z}_1^s := (Z_{a(2s-1)+1}, \ldots, Z_{a(2s-1)+a}) : s \in [\mu]\}.$$

Both $\mathbf{Z}_0$ and $\mathbf{Z}_1$ are made of μ blocks of a consecutive random variables. The blocks are interdependent as well as the variables within each block. $\mathbf{D}_0$ will denote the distribution of $\mathbf{Z}_0$.

We now define a sequence $\underline{\mathbf{Z}}$ of independent blocks as:

$$\underline{\mathbf{Z}} := \{\underline{\mathbf{Z}}^s := (Z_1^s, \ldots, Z_a^s) : s \in [\mu]\},$$

such that the blocks $\underline{\mathbf{Z}}^s$ are mutually independent and such that each block $\underline{\mathbf{Z}}^s$ has the same distribution as $\mathbf{Z}_0^s$, that is, from the stationarity assumption, the distribution of $\mathbf{Z}_0^1$ (the blocks $\underline{\mathbf{Z}}^s$ are IID).

The dependency graph $\underline{\Gamma}$ of $\underline{\mathbf{Z}}$ is such that all the variables in a block are all connected and such that there are no connections between blocks. Theorem 8 can readily be applied to the random sample $\underline{\mathbf{Z}}$, whose distribution we denote $\underline{\mathbf{D}}$: for all P and $\delta \in (0, 1]$,

$$\mathbb{P}_{\underline{\mathbf{Z}} \sim \underline{\mathbf{D}}} \left(\Phi(P, \underline{\mathbf{Z}}, \delta)\right) < \delta, \tag{20}$$

with $e_Q := \mathbb{E}_{\underline{\mathbf{Z}} \sim \underline{\mathbf{D}}} \hat{e}_Q(\underline{\mathbf{Z}})$ and $\Phi(P, \mathbf{Z}, \delta)$ is the event defined as:

$$\Phi(P, \mathbf{Z}, \delta) := \left\{ \exists Q,\ \mathrm{kl}(\hat{e}_Q(\mathbf{Z}) || e_Q) > \frac{1}{\mu} \left[\mathrm{KL}(Q||P) + \ln \frac{\mu + 1}{\delta} \right] \right\}.$$

To see why and how Theorem 8 can be used to get (20), observe that:

- the number of variables in $\underline{\mathbf{Z}}$ is μa;
- by stationarity, all variables Z_α^s, for $\alpha \in [a]$ and $s \in [\mu]$ share the same distribution: we therefore do actually work with dependent but identically distributed variables;
- the (fractional) chromatic number $\underline{\chi}^*$ of $\underline{\Gamma}$ is a, since
  1. the clique number is a (i.e. the number of variables in each block),
  2. the cover $\underline{\mathbf{C}}$ of $\underline{\Gamma}$ with

$$\underline{\mathbf{C}} := \left\{ \left(C_\alpha := \{Z_\alpha^1, \ldots, Z_\alpha^\mu\}, 1\right) \right\}_{1 \leq \alpha \leq a}$$

  is a proper exact cover of size a.

Noting that, consequently

$$\frac{\underline{\chi}^*}{\mu a} = \frac{a}{\mu a} = \frac{1}{\mu} \qquad \text{and} \qquad \frac{\mu a + \underline{\chi}^*}{\delta \underline{\chi}^*} = \frac{\mu a + a}{\delta a} = \frac{\mu + 1}{\delta}$$

gives the expression of $\Phi(P, \mathbf{Z}, \delta)$ and (20).

The last two steps of the proof are similar to those used by Mohri and Rostamizadeh (2009) to establish their bound.

A bound for $\mathbf{Z}_0$. To establish the bound for $\mathbf{Z}_0$, it suffices to use Corollary 24 (Appendix) with $c(\mathbf{z})$ being defined as:

$$c(\mathbf{z}) := \mathbb{I}_{\Phi(P,\mathbf{z},\delta)},$$

which is a bounded measurable function on the blocks $\mathbf{Z}_0^s$ (and thus on the blocks $\underline{\mathbf{Z}}_s$). We have:

$$\left|\mathbb{E}_{\mathbf{Z}_0\sim\mathbf{D}_0} c(\mathbf{Z}_0) - \mathbb{E}_{\underline{\mathbf{Z}}\sim\underline{\mathbf{D}}} c(\underline{\mathbf{Z}})\right| \leq (\mu-1)\beta(a),$$

and therefore, since $\mathbb{P}_{\mathbf{Z}_0\sim\mathbf{D}_0}(\Phi(P,\mathbf{Z}_0,\delta)) = \mathbb{E}_{\mathbf{Z}_0\sim\mathbf{D}_0} c(\mathbf{Z}_0)$ and $\mathbb{P}_{\underline{\mathbf{Z}}\sim\underline{\mathbf{D}}}(\Phi(P,\underline{\mathbf{Z}},\delta)) = \mathbb{E}_{\underline{\mathbf{Z}}\sim\underline{\mathbf{D}}} c(\underline{\mathbf{Z}})$:

$$\begin{aligned}
\mathbb{P}_{\mathbf{Z}_0\sim\mathbf{D}_0}(\Phi(P,\mathbf{Z}_0,\delta)) &\leq \mathbb{P}_{\underline{\mathbf{Z}}\sim\underline{\mathbf{D}}}(\Phi(P,\underline{\mathbf{Z}},\delta)) + (\mu-1)\beta(a) && (21)\\
&< \delta + (\mu-1)\beta(a). && (\text{cf. } (20))
\end{aligned}$$

Establishing the bound. Finally, observe that:

$$\begin{aligned}
\Phi(P,\mathbf{Z},\delta) &\Rightarrow \exists Q:\ \frac{1}{2}\,\mathrm{kl}(\hat{e}_Q(\mathbf{Z}_0)||e_Q) + \frac{1}{2}\,\mathrm{kl}(\hat{e}_Q(\mathbf{Z}_1)||e_Q) > \frac{1}{\mu}\left[\mathrm{KL}(Q||P) + \ln\frac{\mu+1}{\delta}\right]\\
&\Rightarrow \exists Q:\ \bigvee_{i\in\{0,1\}}\left\{\mathrm{kl}(\hat{e}_Q(\mathbf{Z}_i)||e_Q)) > \frac{1}{\mu}\left[\mathrm{KL}(Q||P) + \ln\frac{\mu+1}{\delta}\right]\right\}\\
&\Rightarrow \bigvee_{i\in\{0,1\}}\left\{\exists Q:\mathrm{kl}(\hat{e}_Q(\mathbf{Z}_i)||e_Q)) > \frac{1}{\mu}\left[\mathrm{KL}(Q||P) + \ln\frac{\mu+1}{\delta}\right]\right\}\\
&\Leftrightarrow \Phi(P,\mathbf{Z}_0,\delta)\vee\Phi(P,\mathbf{Z}_1,\delta),
\end{aligned}$$

where we used $\hat{e}_Q(\mathbf{Z}) = \hat{e}_Q(\mathbf{Z}_0)/2 + \hat{e}_Q(\mathbf{Z}_1)/2$ and the convexity of kl in the first line.

This leads to:

$$\begin{aligned}
\mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}(\Phi(P,\mathbf{Z},\delta)) &\leq \mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}(\Phi(P,\mathbf{Z}_0,\delta)\vee\Phi(P,\mathbf{Z}_1,\delta))\\
&\leq \mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}(\Phi(P,\mathbf{Z}_0,\delta)) + \mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}(\Phi(P,\mathbf{Z}_1,\delta)) && (\text{union bound})\\
&= 2\mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m^\beta}(\Phi(P,\mathbf{Z}_0,\delta)) && (\text{stationarity})\\
&= 2\mathbb{P}_{\mathbf{Z}_0\sim\mathbf{D}_0}(\Phi(P,\mathbf{Z}_0,\delta)) && (\text{marginalization wrt } \mathbf{Z}_0)\\
&\leq 2\delta + 2(\mu-1)\beta(a). && (\text{cf. } (21))
\end{aligned}$$

Adjusting δ to $\delta/2 - (\mu-1)\beta(a)$ ends the proof. ■

5. Conclusion

In this work, we propose the first Pac-Bayes bounds applying for classifiers trained on non-IID data. The derivation of these results rely on the use of fractional covers of graphs, convexity and standard tools from probability theory. The results that we provide are very general and can easily be instantiated for specific learning settings such as ranking and learning from from mixing distributions: amazingly, we obtain at a very low cost original Pac-Bayes bounds for these settings. Using a generalized Pac-Bayes bound, we provide

in the appendix a chromatic PAC-Bayes bound that holds for non-independently and non-identically distributed data: it allows us to derive a PAC-Bayes bound for classifiers trained on data from a stationary φ-mixing distribution.

This work gives rise to many interesting questions. First, it seems that using a fractional cover to decompose the non-IID training data into sets of IID data and then tightening the bound through the use of the chromatic number is some form of variational relaxation as often encountered in the context of inference in graphical models, the graphical model under consideration in this work being one that encodes the dependencies in $\mathbf{D}_m$. It might be interesting to make this connection clearer to see if, for instance, tighter and still general bounds can be obtained with more appropriate variational relaxations than the one incurred by the use of fractional covers.

Besides, Theorem 5 advocates for the learning algorithm described in Remark 7. We would like to see how such a learning algorithm based on possibly multiple priors/multiple posteriors could perform empirically and how tight the proposed bound could be.

On another empirical side, it might be interesting to run simulations on bipartite ranking problems to see how accurate the bound of Theorem 15 can be: we expect the results to be of good quality, because of the resemblance of the bound of the theorem with the IID PAC-Bayes theorem for margin classifiers, which has proven to be rather accurate Langford (2005). The work of Germain et al. (2009) is also another contribution that tends to support that a practical use of our bounds should provide competitive results (note that Theorem 25 gives a sufficient condition for the general PAC-Bayes bound of Germain et al. (2009) to be non degenerate). Likewise, it would be interesting to see how the possibly more accurate PAC-Bayes bound for large margin classifiers proposed by Langford and Shawe-taylor (2002), which should translate to the case of bipartite ranking as well, performs empirically. The question also remains as to what kind of strategies to learn the prior(s) could be used to render the bound of Theorem 5 the tightest possible. This is one of the most stimulating question as performing such prior learning makes it possible to obtain very accurate generalization bound Ambroladze et al. (2007).

The connection between our ranking bounds and the theory of U-statistics makes it possible to envision the use of higher order moments in establishing PAC-Bayes bounds, thanks to Hoeffding's decomposition. We plan to investigate further in this direction, for both the ranking measures we have studied (noting that the AUC is a two-sample U-statistics (Hoeffding, 1963)).

Finally, we have been working on a more general way to establish chromatic bounds from IID bounds (covering VC, Rademacher, PAC-Bayes and – possibly – binomial tail bounds), without the need to perform 'low-level' calculations such as the ones proposed in section 3.4. The meta-bound that we have been developping is in the spirit of that proposed by Blanchard and Fleuret (2007), except that the randomization we propose is on the subsets constituting the fractional cover (and not the hypothesis set). In other terms, given a cover $\mathbf{C} = \{(C_j, \omega_j)\}_j$, the fact that an IID bound holds on one subset C_j of a cover is considered as a random event, the probability of a subset to be chosen being $\omega_j/\omega(\mathbf{C})$. A simple union bound gives our generic result, which translates into cover-independent (but fractional-chromatic-number-dependent) chromatic bounds such as (6) (Theorem 8) under very mild conditions on the shape of the base IID bound. Along with that work, we try to answer the question of establishing a principled way to handle situations where random variables

show weak dependencies (as is the case for β-mixing processes), as for now, the framework described here applies when variables are either dependent or independent, disregarding the magnitude of the dependencies – our PAC-Bayes bound for β-mixing processes would then be a specific case of such general result.

Acknowledgment

This work is partially supported by the IST Program of the EC, under the FP7 Pascal 2 Network of Excellence, ICT-216886-NOE.

6. Appendix

6.1 Technical Lemmas

Lemma 20 *Let D be a distribution over $\mathcal{Z}$.*

$$\forall h \in \mathcal{H}, \mathbb{E}_{\mathbf{Z}\sim D^m} e^{m\,\mathrm{kl}\left(\hat{R}(h,\mathbf{Z})||R(h)\right)} \leq m+1.$$

Proof Let $h \in \mathcal{H}$. For $\mathbf{z} \in \mathcal{Z}^m$, we let $q(\mathbf{z}) = \hat{R}(h,\mathbf{z})$; we also let $p = R(h)$. Note that since $\mathbf{Z}$ is i.i.d, $mq(\mathbf{Z})$ is binomial with parameters m and p (recall that $r(h,Z)$ takes the values 0 and 1 upon correct and erroneous classification of Z by h, respectively).

$$\begin{aligned}
\mathbb{E}_{\mathbf{Z}\sim D^m} e^{m\,\mathrm{kl}(q(\mathbf{Z})||p)} &= \sum_{\mathbf{z}\in\mathcal{Z}^m} e^{m\,\mathrm{kl}(q(\mathbf{z})||p)} \mathbb{P}_{\mathbf{Z}\sim D^m}(\mathbf{Z}=\mathbf{z}) \\
&= \sum_{0\leq k\leq m} e^{m\,\mathrm{kl}\left(\frac{k}{m}||p\right)} \mathbb{P}_{\mathbf{Z}\sim D^m}(mq(\mathbf{Z})=k) \\
&= \sum_{0\leq k\leq m} \binom{m}{k} e^{m\,\mathrm{kl}\left(\frac{k}{m}||p\right)} p^k (1-p)^{m-k} \\
&= \sum_{0\leq k\leq m} \binom{m}{k} e^{m\left(\frac{k}{m}\ln\frac{k}{m} + (1-\frac{k}{m})\ln(1-\frac{k}{m})\right)} \\
&= \sum_{0\leq k\leq m} \binom{m}{k} \left(\frac{k}{m}\right)^k \left(1-\frac{k}{m}\right)^{m-k}.
\end{aligned}$$

However, it is obvious that, from the definition of the binomial distribution,

$$\forall m \in \mathcal{N}, \forall k \in [0,m], \forall t \in [0,1], \binom{m}{k} t^k (1-t)^{m-k} \leq 1.$$

This is obviously the case for $t = \frac{k}{m}$, which gives

$$\sum_{0\leq k\leq m} \binom{m}{k} \left(\frac{k}{m}\right)^k \left(1-\frac{k}{m}\right)^{m-k} \leq \sum_{0\leq k\leq m} 1 = m+1.$$

■

Theorem 21 (Jensen's inequality) *Let $f \in \mathcal{R}^{\mathcal{X}}$ be a convex function. For all probability distribution P on $\mathcal{X}$:*

$$f(\mathbb{E}_{X\sim P} X) \leq \mathbb{E}_{X\sim P} f(X).$$

Theorem 22 (Markov's Inequality) *Let X be a positive random variable on $\mathcal{R}$, such that $\mathbb{E}X < \infty$.*

$$\forall t \in \mathcal{R}, \mathbb{P}_X \left\{ X \geq \frac{\mathbb{E}X}{t} \right\} \leq \frac{1}{t}.$$

Consequently: $\forall M \geq \mathbb{E}X, \forall t \in \mathcal{R}, \mathbb{P}_X \left\{ X \geq \frac{M}{t} \right\} \leq \frac{1}{t}$.

Lemma 23 (Convexity of kl) $\forall p, q, r, s \in [0,1], \forall \alpha \in [0,1]$,

$$\mathrm{kl}(\alpha p + (1-\alpha)q || \alpha r + (1-\alpha)s) \leq \alpha\, \mathrm{kl}(p||r) + (1-\alpha)\, \mathrm{kl}(q||s).$$

Proof It suffices to see that $f \in \mathcal{R}^{[0,1]^2}, f(\mathbf{v} = [p\ q]) = \mathrm{kl}(q||p)$ is convex over $[0,1]^2$: the Hessian H of f is

$$H = \begin{bmatrix} \frac{q}{p^2} + \frac{1-q}{(1-p)^2} & -\frac{1}{p} - \frac{1}{1-p} \\ -\frac{1}{p} - \frac{1}{1-p} & \frac{1}{q} + \frac{1}{1-q} \end{bmatrix},$$

and, for $p, q \in [0,1]$, $\frac{q}{p^2} + \frac{1-q}{(1-p)^2} \geq 0$ and $\det H = \frac{(p-q)^2}{q(1-q)p^2(1-p)^2} \geq 0$: $H \succeq 0$ and f is indeed convex. ■

Finally, we have the following version by Mohri and Rostamizadeh (2009) of Corollary 2.7 in (Yu, 1994), which is based on the definition of the blocks $\mathbf{Z}_k^s$:

Corollary 24 *Let c be a measurable function defined with respect to the blocks $\mathbf{Z}_0^s$. If c has absolute value bounded by M, then*

$$|\mathbb{E}_{\mathbf{Z}_0\sim\mathbf{D}_0} c(\mathbf{Z}) - \mathbb{E}_{\underline{\mathbf{Z}}\sim\underline{\mathbf{D}}} c(\underline{\mathbf{Z}})| \leq (\mu - 1) M \beta(a).$$

6.2 Applications of a Generic Pac-Bayes Theorem

Let us first recall the following generic Pac-Bayes result, which is a corollary/compound of results proposed by Seeger (2002b) and McAllester (2003). In particular, the γ function need not be differentiable with respect to its second argument and it applies to any 'risk' functional ψ for which a concentration inequality exists.

Corollary 25 (Generic Pac-Bayes Theorem) *Let $\mathcal{H} \subseteq \mathcal{R}^{\mathcal{X}}$ and $\psi : \mathcal{H} \times \bigcup_{m=1}^{\infty} \mathcal{Z}^m \to \mathcal{R}$. If there exist $\alpha \geq 1, \beta > 1$ and a nonnegative convex function $\Delta : \mathcal{R} \times \mathcal{R} \to \mathcal{R}_+$ that is strictly increasing with respect to its second argument such that*

$$\forall h \in \mathcal{H}, \forall \varepsilon > 0,\ \mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m} \left[\mathbb{E}\psi(h) - \psi(h, \mathbf{Z}) \geq \varepsilon \right] \leq \alpha \exp\left(-\beta \Delta(\mathbb{E}\psi(h), \varepsilon) \right), \tag{22}$$

where $\mathbb{E}\psi(h)$ stands for $\mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \psi(h, \mathbf{Z})$, then, $\forall P$, with probability at least $1-\delta$ over the draw of $\mathbf{Z} \sim \mathbf{D}_m$:

$$\forall Q,\ \Delta(e_Q^{\psi}, e_Q^{\psi} - \hat{e}_Q^{\psi}(\mathbf{Z})) \leq \frac{1}{\beta - 1} \left[\mathrm{KL}(Q||P) + \ln \frac{\alpha\beta}{\delta} \right]. \tag{23}$$

where

$$\hat{e}_Q^{\psi}(\mathbf{Z}) := \mathbb{E}_{h\sim Q}\psi(h,\mathbf{Z})$$
$$e_Q^{\psi} := \mathbb{E}_{\mathbf{Z}}\hat{e}_Q^{\psi}(\mathbf{Z}) = \mathbb{E}_{h\sim Q}\mathbb{E}_{\mathbf{Z}}\psi(h,\mathbf{Z})$$

Proof Along lines from (Seeger, 2002b) and (McAllester, 2003).

1. Observe that, thanks to Lemma 26 (below) with $\delta(\varepsilon) := \Delta(\mathbb{E}\psi(h),\varepsilon)$,

   $$\mathbb{E}_{\mathbf{Z}}e^{(\beta-1)\Delta(\mathbb{E}\psi(h),\mathbb{E}\psi(h)-\psi(h,\mathbf{Z}))} \leq \alpha\beta, \text{ and, } \mathbb{E}_{h\sim P}\mathbb{E}_{\mathbf{Z}}e^{(\beta-1)\Delta(\mathbb{E}\psi(h),\mathbb{E}\psi(h)-\psi(h,\mathbf{Z}))} \leq \alpha\beta$$

   Applying Markov's inequality then gives:

   $$\mathbb{P}_{\mathbf{Z}}\left[\mathbb{E}_{h\sim P}e^{(\beta-1)\Delta(\mathbb{E}\psi(h),\mathbb{E}\psi(h)-\psi(h,\mathbf{Z}))} \geq \frac{\alpha\beta}{\delta}\right] \leq \delta$$

2. Using the entropy extremal inequality $\ln \mathbb{E}_{X\sim P}f(X) \geq -\operatorname{KL}(Q||P)) + \mathbb{E}_{X\sim Q}\ln f(X)$, $\forall P, Q, X$ (see the proof of Lemma 11), and the fact that $x \mapsto \ln x$ is nondecreasing, the previous step leads to

   $$\mathbb{P}_{\mathbf{Z}}\left[\exists Q : -\operatorname{KL}(Q||P) + (\beta-1)\mathbb{E}_{h\sim Q}\Delta(\mathbb{E}\psi(h),\mathbb{E}\psi(h)-\psi(h,\mathbf{Z})) \geq \ln\frac{\alpha\beta}{\delta}\right] \leq \delta.$$

3. Since Δ is convex, Jensen's inequality can be used to give (here, $h \sim Q$)

   $$\mathbb{P}_{\mathbf{Z}}\left[\exists Q : -\operatorname{KL}(Q||P) + (\beta-1)\Delta(\mathbb{E}_{h,\mathbf{Z}}\psi(h,\mathbf{Z}),\mathbb{E}_{h,\mathbf{Z}}\psi(h,\mathbf{Z}) - \mathbb{E}_h\psi(h,\mathbf{Z})) \geq \ln\frac{\alpha\beta}{\delta}\right] \leq \delta.$$

■

Lemma 26 (McAllester (2003)) *Let X be a real-valued random variable on $\mathcal{X}$ and $\alpha \geq 1, \beta > 1$. Let $\delta : \mathcal{R} \to \mathcal{R}$ be a nonnegative and strictly increasing function. We have:*

$$\forall x \in \mathcal{R},\ \mathbb{P}[X \geq x] \leq \alpha e^{-\beta\delta(x)} \Rightarrow \mathbb{E}\left[e^{(\beta-1)\delta(X)}\right] \leq \alpha\beta.$$

Proof See the proof of McAllester (2003). Here, we take α into account. As f is strictly increasing:

$$\mathbb{P}\left[X \geq x\right] = \mathbb{P}\left[\delta(X) \geq \delta(x)\right] = \mathbb{P}\left[e^{(\beta-1)\delta(X)} \geq e^{(\beta-1)\delta(x)}\right].$$

Hence: $\mathbb{P}\left[e^{(\beta-1)\delta(X)} \geq e^{(\beta-1)\delta(x)}\right] \leq \alpha e^{-\beta\delta(x)}$. Setting $\nu = e^{(\beta-1)\delta(x)}$, we get:

$$\mathbb{P}\left[e^{(\beta-1)\delta(X)} \geq \nu\right] \leq \min(1, \alpha\nu^{-\beta/(\beta-1)}).$$

Thus, as for a nonnegative random variable W, $\mathbb{E}[W] = \int_0^\infty \mathbb{P}[W \geq \nu]d\nu$:

$$\mathbb{E}\left[e^{(\beta-1)\delta(X)}\right] \leq 1 + \alpha\int_1^\infty \nu^{-\beta/(\beta-1)} = 1 + \alpha(\beta-1).$$

Since $\alpha > 1$, $1 + \alpha(\beta-1) \leq \alpha\beta$, which ends the proof. ■

We observe that:

- if $\psi(h, \mathbf{Z}) = \sum_{i=1}^{m} \mathbb{I}_{Y_i h(Xi)<0}$ then, by the one-sided Chernoff bound, $\alpha = 1$, $\beta = m$ and $\Delta(p, \varepsilon) = \mathrm{kl}(p - \varepsilon||p)$ make equation (22) hold. The PAC-Bayes bound provided by Corollary 25 is that of Theorem 1 where m is replaced by $m - 1$;
- if

$$\forall i \in [m], \sup_{z_1,\ldots,z_m,z_i' \in \mathcal{Z}} |\psi(z_1, \ldots, z_m) - \psi(z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_m)| \leq c_i,$$

  then, thanks to McDiarmid inequality (McDiarmid, 1989), $\alpha = 1$, $\beta = 2/\sum_i c_i^2$ and $\Delta(p, \varepsilon) = \varepsilon^2$, make equation (22) hold and a PAC-Bayes bound can be derived (we let the reader write the corresponding PAC-Bayes bound);
- it suffices to have an appropriate concentration inequality for the problem at hand to have an effective PAC-Bayes bound.

6.2.1 Generalized Chromatic PAC-Bayes Bound

To get a chromatic PAC-Bayes theorem for non-identically non-independently distributed data, we simply make use of the following concentration inequality of Janson (2004).

Theorem 27 (Janson (2004)) *Suppose that $\mathbf{Z} = \{Z_i\}_{i=1}^m$ is an m-sample of real-valued random variables distributed according to some distribution $\mathbf{D}_m$. Suppose that each Z_i has range $[a_i, b_i]$. If $S_{\mathbf{Z}} = \sum_{i=1}^m Z_i$, then,*

$$\forall \varepsilon > 0,\ \mathbb{P}_{S_{\mathbf{Z}}} \left[\mathbb{E}S_{\mathbf{Z}} - S_{\mathbf{Z}} \geq \varepsilon\right] \leq \exp\left[-\frac{2\varepsilon^2}{\chi^*(\mathbf{D}_m) \sum_{i=1}^m (b_i - a_i)^2}\right],$$

where $\chi^*(\mathbf{D}_m)$ is the fractional chromatic number of the dependency graph of $\mathbf{D}_m$.

Note that *no assumption* is made on the Z_i's being identically distributed.

This concentration inequality gives rise to the following generalized chromatic PAC-Bayes bound that applies to non indepently, possibly non identically distributed data and allows us to use any bounded loss functions r.

Theorem 28 (Generalized Chromatic PAC-Bayes Bound) *$\forall \mathbf{D}_m$, $\forall \mathcal{H}$, $\forall \delta \in (0, 1]$, $\forall P$, with probability at least $1 - \delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m$, the following holds*

$$\forall Q,\ |\hat{e}_Q(\mathbf{Z}) - e_Q|^2 \leq \frac{\chi^* M^2}{2m - \chi^* M^2} \left[\mathrm{KL}(Q||P) + \ln \frac{2m}{\chi^* M^2} + \ln \frac{1}{\delta}\right], \tag{24}$$

where χ^* stands for $\chi^*(\mathbf{D}_m)$, r is a bounded function with range M and

$$\hat{e}_Q(\mathbf{Z}) := \mathbb{E}_{h\sim Q} \hat{R}(h, \mathbf{Z})$$
$$e_Q := \mathbb{E}_{h\sim Q} \hat{e}_Q(\mathbf{Z}) = \mathbb{E}_{h\sim Q} \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m} \hat{R}(h, \mathbf{Z}),$$

with $\hat{R}(h, \mathbf{Z}) := \sum_i r(h, Z_i)/m$.

Proof It suffices to apply Corollary 25 with Theorem 27, $\alpha = 1$, $\Delta(p, \varepsilon) = \varepsilon^2$ and $\beta = 2m/\chi^* M$ (since, as r has range M, $\hat{R}$ has range M/m). ■

We notice the following.

- Here, as no assumption is done regarding the identical distribution of the Z_i's, the expected risk $R(h) = \mathbb{E}_{\mathbf{Z}}\hat{R}(h, \mathbf{Z})$ does not unfold as in (3).

- In the case of using identically distributed random variables and the 0-1 loss, there is no concentration inequality that allows us to retrieve the tighter Pac-Bayes bound given in Theorem 8.

- From a more general point of view, it is enticing to try to establish even more generic results resting on the principle of graph coloring with the aim of decoupling this approach to the PAC-Bayesian framework. This is the subject of ongoing work.

6.2.2 φ-mixing Pac-Bayes Bound

The definition of a φ-mixing process follows.

Definition 29 (φ-mixing process) *Let $\mathbf{Z} = \{Z_t\}_{t=-\infty}^{+\infty}$ be a stationary sequence of random variables. For any $i, j \in \mathbb{Z} \cup \{-\infty, +\infty\}$, let σ_i^j denote the σ-algebra generated by the random variables Z_k, $i \leq k \leq j$. Then, for any positive integer k, the φ-mixing coefficient $\varphi(k)$ of the stochastic process $\mathbf{Z}$ is defined as*

$$\varphi(k) = \sup_{n, A \in \sigma_{n+k}^{+\infty}, B \in \sigma_{-\infty}^{n}} |\mathbb{P}[A|B] - \mathbb{P}[A]|. \tag{25}$$

$\mathbf{Z}$ is said to be φ-mixing if $\varphi(k) \to 0$ as $k \to 0$.

In order to establish our new Pac-Bayes bounds for stationary φ-mixing distributions, it suffices to make use of the following concentration inequality by Kontorovich and Ramanan (2008).

Theorem 30 (Kontorovich and Ramanan (2008)) *Let $\psi : \mathcal{U}^m \to \mathcal{R}$ be a function defined over a countable space $\mathcal{U}$. If ψ is l-Lipschitz with respect to the Hamming metric for some $l > 0$, then the following holds for all $t > 0$:*

$$\mathbb{P}_{\mathbf{Z}}[|\psi(\mathbf{Z}) - \mathbb{E}_{\mathbf{Z}}[\psi(\mathbf{Z})]| > t] \leq 2\exp\left[-\frac{t^2}{2ml^2\|\Lambda_m\|_\infty^2}\right],$$

where $\|\Lambda_m\|_\infty \leq 1 + 2\sum_{k=1}^{m}\varphi(k)$.

Suppose that the loss function r is again such that it takes values in $[0, M]$. Then, for any $h \in \mathcal{H}$, the function $\psi(\mathbf{Z}) = \frac{1}{m}\sum_{i=1}^{m} r(h, Z_i) = \hat{R}(h, \mathbf{Z})$ is obviously M/m-Lipschitz. Therefore, for a sample $\mathbf{Z}$ drawn according to a φ-mixing process, we have the following concentration inequality on $\hat{R}(h, \mathbf{Z})$ that holds for any $h \in \mathcal{H}$:

$$\mathbb{P}_{\mathbf{Z}\sim\mathbf{D}_m}\left[\left|\hat{R}(h, \mathbf{Z}) - R(h)\right| > t\right] \leq 2\exp\left[-\frac{mt^2}{2M^2\|\Lambda_m\|_\infty^2}\right]. \tag{26}$$

We directly get the following Pac-Bayes bound for φ-mixing processes.

Theorem 31 (PAC-Bayes bound for stationary φ-mixing processes) *Let $\mathbf{D}^{\varphi}$ be a stationary φ-mixing distribution over $\mathcal{Z}$ and $\mathbf{D}_m^{\varphi}$ be the distribution of m-samples according to $\mathbf{D}^{\varphi}$. $\forall \mathcal{H} \subseteq \mathcal{R}^{\mathcal{X}}$, $\forall \delta \in (0,1]$, $\forall P$, with probability at least $1-\delta$ over the random draw of $\mathbf{Z} \sim \mathbf{D}_m^{\varphi}$, the following holds*

$$\forall Q,\ |\hat{e}_Q^{\varphi}(\mathbf{Z}) - e_Q^{\varphi}|^2 \leq \frac{2M^2\|\Lambda_m\|_\infty^2}{m - 2M^2\|\Lambda_m\|_\infty^2}\left[\mathrm{KL}(Q||P) + \ln\frac{m}{M^2\|\Lambda_m\|_\infty^2} + \ln\frac{1}{\delta}\right],$$

where $\|\Lambda_m\|_\infty \leq 1 + 2\sum_{k=1}^{m}\varphi(k)$, $r(h,Z) = \mathbb{I}_{Yh(X)<0}$ and

$$\hat{e}_Q^{\varphi}(\mathbf{Z}) := \mathbb{E}_{h\sim Q}\hat{R}(h,\mathbf{Z}) = \mathbb{E}_{h\sim Q}\sum_{t=1}^{m}\mathbb{I}_{Y_t h(X_t)<0}$$
$$e_Q^{\varphi} := \mathbb{E}_{\mathbf{Z}\sim\mathbf{D}_m^{\varphi}}\hat{e}_Q^{\varphi}(\mathbf{Z}).$$

Proof Equation (26), and Corollary 25 with $\alpha = 2$, $\beta = m/(2M^2\|\Lambda\|_\infty^2)$, $\Delta(p,\varepsilon) = \varepsilon^2$. ■

References

S. Agarwal, T. Graepel, R. Herbrich, S. Har-Peled, and D. Roth. Generalization Bounds for the Area Under the ROC Curve. *Journal of Machine Learning Research*, 6:393–425, 2005.

Shivani Agarwal and Partha Niyogi. Generalization bounds for ranking algorithms via algorithmic stability. *Journal of Machine Learning Research*, 10:441–474, 2009.

A. Ambroladze, E. Parrado-Hernandez, and J. Shawe-Taylor. Tighter PAC-Bayes Bounds. In *Adv. in Neural Information Processing Systems 19*, pages 9–16, 2007.

K. Ataman, W. Nick, and Y. Zhang. Learning to rank by maximizing auc with linear programming. In *In IEEE International Joint Conference on Neural Networks (IJCNN 2006)*, pages 123–129, 2006.

J.-Y. Audibert and O. Bousquet. Combining PAC-Bayesian and Generic Chaining Bounds. *Journal of Machine Learning Research*, 8:863–889, 2007. ISSN 1533-7928.

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

P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

G. Blanchard and F. Fleuret. Occam's hammer. In *COLT*, pages 112–126, 2007.

O. Bousquet and A. Elisseeff. Stability and Generalization. *Journal of Machine Learning Research*, 2:499–526, March 2002.

U. Brefeld and T. Scheffer. AUC Maximizing Support Vector Learning. In *Proc. of the ICML Workshop on ROC Analysis in Machine Learning*, 2005.

O. Catoni. *Pac-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, volume 56 of *Lecture Notes–Monograph Series*. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2007.

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

C. Cortes and M. Mohri. AUC Optimization vs. Error Rate Minimization. In *Adv. in Neural Information Processing Systems 16*, 2004.

Y. Freund, R. Iyer, R.E. Schapire, and Y. Singer. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, 4:933–969, 2003.

P. Germain, A. Lacasse, F. Laviolette, and M. Marchand. PAC-Bayesian learning of linear classifiers. In *Proc. of the 26th Annual International Conference on Machine Learning*, pages 353–360, 2009.

R. Herbrich and T. Graepel. A pac-bayesian margin bound for linear classifiers: Why svms work. In *Advances in Neural Information Processing Systems 13*, pages 224–230, 2001.

W. Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *Annals of Mathematical Statistics*, 19(3):293–325, 1948.

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

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

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

A. Lacasse, F. Laviolette, M. Marchand, P. Germain, and N. Usunier. PAC-Bayes Bounds for the Risk of the Majority Vote and the Variance of the Gibbs Classifier. In *Advances in Neural Information Processing Systems 19*, pages 769–776, 2006.

J. Langford. Tutorial on Practical Theory for Classification. *Journal of Machine Learning Research*, pages 273–306, 2005.

J. Langford and J. Shawe-taylor. PAC-Bayes and Margins. In *Adv. in Neural Information Processing Systems 15*, pages 439–446, 2002.

D. McAllester. Some PAC-Bayesian Theorems. *Machine Learning*, 37:355–363, 1999.

D. McAllester. Simplified pac-bayesian margin bounds. In *Proc. of the 16th Annual Conference on Computational Learning Theory*, pages 203–215, 2003.

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

M. Mohri and A. Rostamizadeh. Rademacher complexity bounds for non-i.i.d. processes. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1097–1104, 2009.

S. V. Pemmaraju. Equitable coloring extends chernoff-hoeffding bounds. In *RANDOM-APPROX*, pages 285–296, 2001.

A. Rakotomamonjy. Optimizing the Area under the ROC curve with SVMs. In *ROC Analysis in Artificial Intelligence*, pages 71–80, 2004.

E.R. Schreinerman and D.H. Ullman. *Fractional graph theory: A rational approach to the theory of graphs*. Wiley Interscience Series in Discrete Math., 1997.

M. Seeger. PAC-Bayesian generalization bounds for gaussian processes. *Journal of Machine Learning Research*, 3:233–269, 2002a.

M. Seeger. The proof of McAllester's PAC-Bayesian theorem. Technical report, Institute for ANC, Edinburgh, UK, 2002b.

N. Usunier, M.-R. Amini, and P. Gallinari. A Data-dependent Generalisation Error Bound for the AUC. In *Proc. of the ICML Workshop on ROC Analysis in Machine Learning*, 2005.

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

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *Annals of Probability*, 22(1):94–116, 1994.