PAC-Bayesian High Dimensional Bipartite Ranking

Benjamin Guedj* and Sylvain Robbiano†

November 9, 2015

Abstract

This paper is devoted to the bipartite ranking problem, a classical statistical learning task, in a high dimensional setting. We propose a scoring and ranking strategy based on the PAC-Bayesian approach. We consider nonlinear additive scoring functions, and we derive non-asymptotic risk bounds under a sparsity assumption. In particular, oracle inequalities in probability holding under a margin condition assess the performance of our procedure, and prove its minimax optimality. An MCMC-flavored algorithm is proposed to implement our method, along with its behavior on synthetic and real-life datasets.

Keywords: Bipartite Ranking, High Dimension and Sparsity, MCMC, PAC-Bayesian Aggregation, Supervised Statistical Learning.

1 Introduction

The bipartite ranking problem appears in various application domains such as medical diagnostic, information retrieval, signal detection. This supervised learning task consists in building a so-called scoring function that order the (high dimensional) observations in the same fashion as the (unknown) associated labels. In that sense, the global problem of bipartite ranking (ordering a set observations) includes the local classification task (assigning a label to each observation). Indeed, once a proper scoring function is defined, classification amounts to choosing a threshold, assigning data points to either class depending on whether their score is above or below that threshold.

*Modal project-team, Inria, France. benjamin.guedj@inria.fr

†Department of Statistical Science, UCL, United Kingdom. sylvain.robbiano@gmail.com

The quality of a scoring function is usually assessed through its ROC curve (Receiver Operating Characteristic, see Green and Swets, 1966) and it is shown that maximizing this visual tool is equivalent to solving the bipartite ranking problem (see for example, Proposition 6 in Clémençon and Vayatis, 2009). Due to the functional nature of the ROC curve, it is useful to substitute a proxy: maximizing the Area Under the ROC Curve (AUC), or equivalently minimizing the pairwise risk. Following that idea, several classical algorithms in classification have been extended to the case of bipartite ranking, such as Rankboost (Freund et al., 2003) or RankSVM (Rakotomamonjy, 2004). Several authors have considered theoretical aspects of this problem. In Agarwal et al. (2005), the authors investigate the difference between the empirical AUC and the true AUC and produce a concentration inequality assessing that as the number of observations increases, the empirical AUC tends to the true AUC, allowing for empirical risk minimization (ERM) approaches to tackle the bipartite ranking problem. This strategy has been explored by Clémençon et al. (2008). Assuming that the true scoring function is in a Vapnik-Cervonenkis class of finite dimension, combined with a low noise condition, the authors prove that the minimizer of the empirical pairwise risk achieves fast rates of convergence. More recently, the bipartite ranking problem has been tackled from a nonparametric angle and Clémençon and Robbiano (2011) proved that a plug-in estimator of the regression function can attain minimax fast rates of convergence over Hölder class. In order to obtain an adaptive estimator to the low noise and to the Hölder parameters, an aggregation procedure based on exponential weights has been studied by Robbiano (2013) and the author shows adaptive fast rate upper bounds. However, the rates of convergence depend on the dimension of the features space and in many applications the optimal scoring function depends on a small number of the features, suggesting a sparsity assumption. As a matter of fact, the problem of sparse bipartite ranking has been studied by Li et al. (2013), for linear scoring function. In this paper, we consider a more general case by introducing nonparametric scoring functions that can be sparsely decomposed in an additive way with respect to the covariates.

To do so, we design an aggregation strategy which heavily relies on the PAC-Bayesian paradigm (the acronym PAC stands for *Probably Approximately Correct*). In our setting, the PAC-Bayesian approach delivers a random estimator (or its expectation) sampled from a (pseudo) posterior distribution which exponentially penalizes the AUC risk. The PAC-Bayesian theory originates in the two seminal papers Shawe-Taylor and Williamson (1997) and McAllester (1999). The first PAC-Bayesian bounds consisted in data-dependent empirical inequalities for Bayesian-flavored estimators. This strategy has been extensively formalized in the context of classification by Catoni (2004, 2007) and regression by Audibert (2004a,b), Alquier (2006, 2008) and Audibert and Catoni (2010, 2011). PAC-Bayesian techniques have proven useful to study the convergence rates of Bayesian learning methods. More recently, these methods have been studied under the scope of high dimensionality (typically with a sparsity assumption): see for example Dalalyan and Tsybakov (2008, 2012), Alquier and Lounici (2011), Dalalyan and Salmon (2012), Suzuki (2012), Alquier and Biau (2013), and Guedj and Alquier (2013). The main message of these works is that PAC-Bayesian aggregation with a properly chosen prior is able to deal effectively with the sparsity issue in a regression setting under the ℓ^2 loss. The purpose of the present paper is to extend the use of such techniques to the case of bipartite ranking. Note that in a work parallel to ours, Ridgway et al. (2014) also use a PAC-Bayesian machinery for bipartite ranking. Their framework is close to the one developed in this paper, however the authors focus on linear scoring function and produce oracle inequality in expectation. The work presented in the present paper is more general as we consider nonlinear scoring functions and derive non-asymptotic risk bounds in probability.

Our procedure relies on the construction of a high dimensional yet sparse (pseudo) posterior distribution (Gibbs distribution, introduced in Section 2). Most of the aforecited papers proposing PAC-Bayesian strategies rely on Monte Carlo Markov Chain (MCMC) algorithms to sample from this target distribution. However, very few discuss the practical implementation of an MCMC algorithm in the case of (possibly very) high dimensional data, to the notable exception of Alquier and Biau (2013), Guedj and Alquier (2013) (MCMC) and Ridgway et al. (2014) (Sequential Monte Carlo).

We adapt the point of view presented in Guedj and Alquier (2013) and implemented in the R package `pacbpred` (Guedj, 2013), which is inspired by the seminal work of Carlin and Chib (1995) and its later extensions Hans et al. (2007) and Petralias and Dellaportas (2012). The key idea is to define a neighborhood relationship between visited models, promoting local moves of the Markov chain. This approach has the merit of being easily implementable, and adapts well to our will to promote sparse scoring functions. Indeed, the Markov chain will mostly visit low dimensional models, ensuring sparse scoring predictors as outputs. As emphasized in the following, this choice leads to nice performance when a sparse additive representation is a good approximation of the optimal scoring function.

The paper is structured as follows. In Section 2, we introduce the notation and our PAC-Bayesian estimation strategy for the bipartite ranking problem. Section 3 contains the main theoretical results of the paper, in the form of oracle inequalities in probability. Risk bounds are presented to assess the merits of our procedure and exhibit explicit rates of convergence. Section 4 presents our MCMC algorithm to compute our estimator, coupled with numerical results on both synthetic and real-life datasets. Conclusive comments on both theoretical and practical merits of our work are summed up in Section 5. Finally, proofs of the original results claimed in the paper are gathered in Section 6 for the sake of clarity.

2 Notation

Adopting the notation $\mathbf{X} = (X_1, \ldots, X_d)$, we let $(\mathbf{X}, Y)$ be a random variable taking its values in $\mathbb{R}^d \times \{\pm 1\}$. We let $\mathbb{P} = (\mu, \eta)$ denote the distribution of $(\mathbf{X}, Y)$, where μ is the marginal distribution of $\mathbf{X}$, and $\eta(\cdot) = \mathbb{P}[Y = 1 | \mathbf{X} = \cdot]$. Our goal is to solve the bipartite ranking problem, *i.e.*, ordering the features space $\mathbb{R}^d$ to preserve the orders induced by the labels. In other words, our goal is to design an order relationship on $\mathbb{R}^d$ which is consistent with the order on $\{\pm 1\}$: when given a new pair of points $(\mathbf{X}, Y)$ and $(\mathbf{X}', Y')$ drawn from $\mathbb{P}$, order $\mathbf{X}$ then $\mathbf{X}'$ iff $Y < Y'$.

A natural way to build up such an order relation on $\mathbb{R}^d$ is to transport the usual order on the real line onto $\mathbb{R}^d$ through a (measurable) *scoring function* $s : \mathbb{R}^d \to \mathbb{R}$ such that

$$\forall (\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d, \qquad \mathbf{x} \preceq_s \mathbf{x}' \Leftrightarrow s(\mathbf{x}) \leq s(\mathbf{x}').$$

It is tempting to try and mimic the sorting performance of the unknown regression function η, which is clearly optimal. The bipartite ranking problem may now be rephrased as building a scoring function s such that, for any pair $(\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d$, $s(\mathbf{x}) \leq s(\mathbf{x}') \Leftrightarrow \eta(\mathbf{x}) \leq \eta(\mathbf{x}')$. From a statistical perspective, our goal is to learn such a scoring function using a n-sample $\mathcal{D}_n = \{(\mathbf{X}_i, Y_i)\}_{i=1}^n$ consisting in i.i.d. replications of $(\mathbf{X}, Y)$.

To assess the theoretical quality of a scoring function, it is natural to consider the ranking risk, based on the pairwise classification loss, defined as follows: let $(\mathbf{X}, Y)$ and $(\mathbf{X}', Y')$ be two independent variables drawn from some distribution $\mathbb{P}$. The ranking risk of some scoring function s is

$$L(s) = \mathbb{P}\left[(s(\mathbf{X}) - s(\mathbf{X}'))(Y' - Y) < 0\right].$$

This quantity is closely related to the AUC,

$$\mathrm{AUC}(s) = \mathbb{P}[s(\mathbf{X}) < s(\mathbf{X}') | Y = -1, Y' = +1],$$

as pointed out by Clémençon et al. (2008). Note that the authors have proved that the optimal scoring function for the ranking risk is the posterior distribution η, which is obviously unknown to the statistician.

The approach adopted in this paper consists in providing bounds on the *excess ranking risk*, defined as

$$\mathcal{E}(\cdot) = L(\cdot) - L^\star,$$

where $L^\star := L(\eta)$. Since η is the minimizer of L, $\mathcal{E}$ is a positive valued function. In Clémençon et al. (2008), it is shown that the excess ranking risk may be reformulated as

$$\mathcal{E}(s) = \mathbb{E}\left[\left|\eta(\mathbf{X}) - \eta(\mathbf{X}')\right| \mathbb{1}_{\{(s(\mathbf{X})-s(\mathbf{X}'))(\eta(\mathbf{X}')-\eta(\mathbf{X}))<0\}}\right], \quad \forall s.$$

The type of bounds we are interested in consists in oracle inequalities in probability, which will depend on the considered family of scoring functions.

Since $\mathbb{P}$ is unknown, the minimizer η of L is unavailable. Instead, we substitute to L its empirical counterpart, the empirical ranking risk L_n defined as

$$L_n : s \mapsto \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{1}_{\{(Y_i - Y_j)(s(\mathbf{X}_i)-s(\mathbf{X}_j))<0\}},$$

and likewise, we let $\mathcal{E}_n(s) := L_n(s) - L_n(\eta)$ denote the empirical excess ranking risk.

This paper is devoted to the case where η admits a sparse representation, *i.e.*, only a small number $d_0 \ll d$ of covariates is necessary to build efficient prediction procedures. Note that this is also the angle studied by Li et al. (2013) in a parallel work to ours, in a less general setting.

Following Guedj and Alquier (2013), we focus on a sparse additive modeling of the optimal scoring function. Indeed, we will build up an estimate from the family

$$\mathcal{S}_\Theta = \left\{ s_\theta : \mathbf{x} \mapsto \sum_{j=1}^{d} \sum_{k=1}^{M} \theta_{jk} \phi_k(x_j), \quad \theta \in \mathbb{R}^{dM} \right\},$$

where $\mathbb{D} = \{\phi_1, \ldots, \phi_M\}$ is a dictionary of deterministic known functions, and we adopt the notation

$$\begin{aligned} \theta &= (\theta_{jk})_{j=1,\ldots,d}^{k=1,\ldots,M} \\ &= (\theta_{11}, \theta_{12}, \ldots, \theta_{1M}, \theta_{21}, \ldots, \theta_{2M}, \ldots, \theta_{dM}). \end{aligned}$$

Our choice for this additive formulation is motivated by the nice compromise achieved between flexibility and interpretation (see for example Hastie and Tibshirani, 1986; Stone, 1985). Our aim is to produce a sparse estimate $\hat{\theta}$ and then compute the plugin estimator $s_{\hat{\theta}}$. To do so, we rely on the PAC-Bayesian approach and we will specify in the following section a sparsity-promoting so-called prior π on Θ embedded with its Borel σ-algebra. Finally, let $\mathbf{m} = (m_1, \dots, m_d) \in \{0,1\}^d$ encode a model (where $m_j = 1$ iff covariate j is present).

From the prior π, we let $\hat{\rho}_\delta$ denote the Gibbs (pseudo-)posterior density, defined as

$$\hat{\rho}_\delta(\mathrm{d}\theta) \propto \exp[-\delta L_n(s_\theta)]\pi(\mathrm{d}\theta), \tag{1}$$

where $\delta > 0$ may be seen as an inverse temperature parameter. This density twists the prior mass towards functions s_θ for which $L_n(s_\theta)$ is not too large. Indeed, if π puts more probability mass on sparse vectors, $\hat{\rho}_\delta$ will favor sparse vectors with a small empirical ranking risk, thus meeting our requirements. The Gibbs pseudo-posterior in (1) has attracted a great deal of interest in recent years (under the name exponentially weighted aggregate as in Dalalyan and Tsybakov, 2008; Rigollet and Tsybakov, 2012, among others).

The final estimator is

$$s_{\hat{\theta}} : \mathbf{X} \mapsto \sum_{j=1}^{d} \sum_{k=1}^{M} \hat{\theta}_{jk} \phi_k(X_j), \tag{2}$$

where

$$\hat{\theta} \sim \hat{\rho}_\delta.$$

Note that for the sake of brevity, we will use the notation $\hat{s} = s_{\hat{\theta}}$. We will provide in Section 3 non-asymptotic oracle inequalities to assess the theoretical merits of the estimator $\hat{s}$. Section 4 is devoted to the practical implementation of $\hat{s}$.

3 Oracle inequalities

In this section, we provide the main theoretical results of the paper, consisting in oracle inequalities in probability for the estimator $\hat{s}$ defined in (2). We specify different rates of convergence under several mild assumptions on the distribution $\mathbb{P}$ of $(\mathbf{X}, Y)$. The only tool we need to derive our first results is an exponential inequality on the difference of the excess ranking risk and its empirical counterpart.

Condition 1. *For any inverse temperature parameter $\delta > 0$, and any candidate function s,*

$$\mathbb{E}\exp[\delta(\mathcal{E}_n(s) - \mathcal{E}(s))] \leq \exp(\psi(s)), \tag{3}$$

where ψ may depend on n and δ.

Note that this concentration condition is classical in the PAC-Bayesian literature, and allows for our first result. We let $\mathcal{K}(\mu, \nu)$ denote the Kullback-Leibler divergence between two measures μ and ν, and we let $\mathcal{M}_\pi$ stand for the space of probability measures which are absolutely continuous with respect to π.

Theorem 1. *Assume that Condition 1 holds. Then, for any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \int \frac{\psi(s)}{\delta}\rho(\mathrm{d}s) + \frac{\psi(\hat{s}) + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1 - \varepsilon.$$

This result is in the spirit of classical PAC-Bayesian bounds such as in Catoni (2004). It ensures that the excess risk of our procedure $\hat{s}$ is bounded with high probability by the mean excess risk of any realization of some posterior distribution ρ absolutely continuous with respect to some prior π, up to remaining terms involving the Kullback-Leibler divergence between ρ and π and the right-hand side term from the exponential inequality (3). Note that if the right-hand term of (3) does not depend on the scoring function, *i.e.*, $\psi(s) = \psi$ for any s, Theorem 1 amounts to the inequality

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \frac{2\psi + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1 - \varepsilon.$$

Next, without any further assumption on the distribution $\mathbb{P}$ of $(\mathbf{X}, Y)$, we are able to precise the right-hand side term in (3).

Corollary 1. *For any distribution of the random variables $(\mathbf{X}, Y)$, Condition 1 holds with $\psi(s) = \delta^2/4n$.*
Hence, using Theorem 1 and choosing $\delta = \sqrt{n}$, for any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \frac{1/2 + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\sqrt{n}}\right\}\right] \geq 1 - \varepsilon.$$

The message here is that we obtain the classical slow rate of convergence $\sqrt{n}$ (as achieved, for example, by empirical AUC minimization on a Vapnik-Cervonenkis class) under no assumption whatsoever on $\mathbb{P}$ with the PAC-Bayesian approach.

3.1 Using a sparsity-promoting prior

Our goal is to obtain *sparse* vectors θ, and this constraint is met with the introduction of the following prior π:

$$\pi(\mathrm{d}\theta) \propto \sum_{\mathbf{m}} \binom{d}{|\mathbf{m}|_0}^{-1} \beta^{|\mathbf{m}|_0 M} \mathrm{Unif}_{\mathcal{B}_{\mathbf{m}}}(\theta), \tag{4}$$

where $\beta \in (0,1)$, $|\mathbf{m}|_0 = \sum_{j=1}^d m_j$, and $\mathcal{B}_{\mathbf{m}}$ denotes the ℓ^2-ball in $\mathbb{R}^{|\mathbf{m}|_0}$ of radius 2. This prior may be traced back to Leung and Barron (2006) and serves our purpose: for any $\theta \in \Theta$, its probability mass will be negligible unless its support has a very small dimension, *i.e.*, θ is sparse. Next, we introduce a technical condition required in our scheme.

Condition 2. *There exists c>0, such that*

$$\mathbb{P}[s_\theta(\mathbf{X}) - s_\theta(\mathbf{X}') \geq 0, s_{\theta'}(\mathbf{X}) - s_{\theta'}(\mathbf{X}') \leq 0] \leq c\|\theta - \theta'\|$$

for any θ and $\theta' \in \mathbb{R}^d$ such that $\|\theta\| = \|\theta'\| = 1$.

This assumption is the exact analogous to the density assumption used in Ridgway et al. (2014) and echoes classical technical requirements linked to margin assumptions, as discussed further (see for example Audibert and Tsybakov, 2007).

The use of this sparsity-inducing prior allows us to obtain terms in the right-hand side of the oracle inequalities which depend on the intrinsic dimension of the ranking problem, *i.e.*, the dimension of the sparsest representation s_θ of the optimal scoring function.

Theorem 2. *For any distribution of the random variables* $(\mathbf{X}, Y)$*, Condition 1 holds with $\psi = \delta^2/4n$. Let $\delta = \sqrt{n}$. With the prior π defined as in (4), under Condition 2, we obtain for any $\varepsilon \in (0,1)$,*

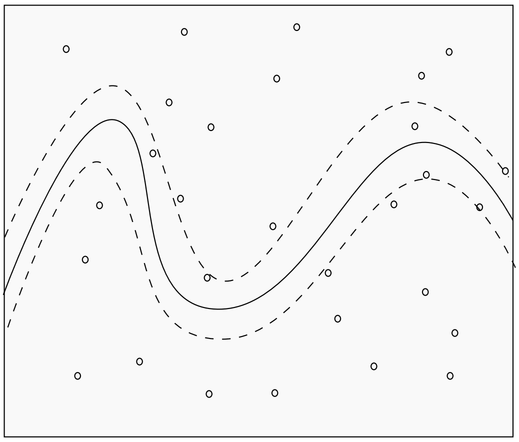
$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}} \inf_{\theta \in \mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \left\{\mathcal{E}(s_\theta) + \frac{3/2 + 2\log(2/\varepsilon) + \log(2c\sqrt{n}) + K}{\sqrt{n}}\right\}\right] \geq 1 - \varepsilon,$$

where

$$K = 2\left(|\mathbf{m}|_0 M \log(1/\beta) + |\mathbf{m}|_0 \log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}\right).$$

This sharp oracle inequality ensures that if there exists indeed a sparse representation (*i.e.*, some sparse model $\mathbf{m}$) of the optimal scoring function, *i.e.*, involving only a small number of covariates, then the excess risk of our procedure is bounded by the best excess risk among all linear combinations of the dictionary up to some small terms. On the contrary, if such a representation does not exist, $|\mathbf{m}|_0$ is comparable to d and terms like $|\mathbf{m}|_0 \log(d)/\sqrt{n}$ and $|\mathbf{m}|_0 M \log(1/\beta)$ start to emerge.

Figure 1 – Margin condition in a simple example: the bold curve is the level set $\eta = t$, $t \in (0,1)$, and the dashed curves are respectively the level sets $\eta = t \pm \epsilon$ for some $\epsilon > 0$. The margin condition ensures that the probability mass of the set in-between the dashed curves is small (with an explicit control, through a certain power of a parameter α), for any value of t.



3.2 Faster rates with a margin condition

In order to obtain faster rates, we follow Robbiano (2013) and work under the following margin condition.

Condition 3. *The distribution of* $(\mathbf{X}, Y)$ *verifies the margin assumption* $\mathbf{MA}(\alpha)$ *with parameter* $0 \le \alpha \le 1$ *if there exists* $C < \infty$ *such that:*

$$\mathbb{P}\left[(s(\mathbf{X}) - s(\mathbf{X}'))(\eta(\mathbf{X}) - \eta(\mathbf{X}')) < 0\right] \le C(L(s) - L^\star)^{\frac{\alpha}{1+\alpha}},$$

for any scoring function s.

This margin (or low noise) condition was first introduced for classification by Mammen and Tsybakov (1999) and extended by Tsybakov (2004), later adapted for the ranking problem by Clémençon et al. (2008). The margin effect is illustrated by Figure 1. Note that this statement is trivial for the value $\alpha = 0$ and increasingly restrictive as α grows. We refer the reader to Boucheron et al. (2005), Lecué (2006) and Robbiano (2013) for an extended discussion.

The next step to our main results is the following lemma.

Lemma 1. *Let* s *be a scoring function and* $(\mathbf{X}, Y)$ *and* $(\mathbf{X}', Y')$ *two pairs of independent random variables. Let* $T(s) = \mathbb{1}_{\{(s(\mathbf{X})-s(\mathbf{X}'))(Y-Y')<0\}} - \mathbb{1}_{\{(\eta(\mathbf{X})-\eta(\mathbf{X}'))(Y-Y')<0\}}$, *and $\mathbb{V}(Z)$ denote the variance of a random variable Z. Let Condition 3 hold for some $\alpha \in (0,1)$, then*

$$\mathbb{V}(T(s)) \le \mathcal{C}(L(s) - L^*)^{\frac{\alpha}{1+\alpha}},$$

where $\mathcal{C}$ is a constant.

Lemma 1 allows us to obtain a tighter right-hand term in (3), therefore leading to a refined oracle inequality with a faster rate of convergence. Let us introduce the notation $\phi(u) = e^u - u - 1$.

Finally, for the sake of brevity, we will use the notation Υ, C_1, C_2 and C_3 for generic constants in the following statements. The exact form of those constants may be found in the proofs (Section 6).

Theorem 3. *For any distribution of the random variables $(\mathbf{X}, Y)$ satisfying Condition 3 for some parameter $\alpha \in (0,1)$, Condition 1 holds with $\psi(s) = \frac{n}{2}\mathbb{V}(T(s))\phi\left(\frac{2\delta}{n}\right)$.*
Hence, using Theorem 1 and choosing $\delta = \Upsilon n^{\frac{1+\alpha}{2+\alpha}}$, for any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\rho \in \mathcal{M}_\pi} \left\{3\int \mathcal{E}(s)\rho(\mathrm{d}s) + n^{-\frac{1+\alpha}{2+\alpha}}\left[C_1 + \Upsilon^{-1}\left(1/2 + \log(2/\varepsilon) + \mathcal{K}(\rho,\pi)\right)\right]\right\}\right] \ge 1-\varepsilon,$$

where Υ and C_1 are constants depending only on c, C, α and $\mathcal{C}$.

Note that taking $\alpha = 0$ yields a rate of convergence similar to the one in Corollary 1. A trade-off is at work in that result: in all generality, the fastest rate achievable is of magnitude $\sqrt{n}$. However, under the margin condition, refined rates are available, at the cost of generality: the greater α, the faster the rate *and* the more restrictive the assumption on $\mathbb{P}$. For more comments on the introduction of margin conditions for ranking problems and its impact on rates of convergence, we refer the reader to the aforecited Clémençon and Robbiano (2011, Section 2.3).

The next result is the adaptation of Theorem 2 under Condition 3, to obtain faster rates.

Theorem 4. *For any distribution of the random variables $(\mathbf{X}, Y)$ satisfying Condition 3 for some parameter $\alpha \in (0,1)$, Condition 1 holds with $\psi = \frac{n}{2}\mathbb{V}(T(s))\phi\left(\frac{2\delta}{n}\right)$.*
Hence, using Theorem 2 and choosing $\delta = \Upsilon n^{\frac{1+\alpha}{2+\alpha}}$, under Condition 2, for any $\varepsilon \in (0,1)$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta \in \mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \left\{3\mathcal{E}(s_\theta) + n^{-\frac{1+\alpha}{2+\alpha}} C_1\left(K + 3/2 + 2\log(2/\varepsilon) + \log\left(n^{\frac{1+\alpha}{2+\alpha}}\right)\right)\right\}\right] \ge 1-\varepsilon,$$

where Υ and C_1 are constants depending only on c, C, α, β and $\mathcal{C}$, and

$$K = 2\left(|\mathbf{m}|_0 M \log(1/\beta) + |\mathbf{m}|_0 \log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}\right).$$

This result walks in the footsteps of previous work on the use of the margin condition in bipartite ranking, such as in Robbiano (2013). As in Theorem 2, Theorem 4 exhibits right-hand terms in the oracle inequality which depend on the intrinsic dimension of the problem, now with a significantly faster rate, of magnitude $n^{\frac{1+\alpha}{2+\alpha}}$.

3.3 Rates of convergence on Sobolev classes

In order to control the bias leading term in the previous oracle inequalities, we refine the previous results under the additional assumption that the regression function now belongs to some functional regularity space. Following Tsybakov (2009), we consider the Sobolev ellipsoid defined as

$$\mathcal{W}(\tau,\kappa) = \left\{ f \in \mathrm{L}^2([-1,1]) \colon f = \sum_{k=1}^{\infty} \theta_k \varphi_k \quad \text{and} \quad \sum_{i=1}^{\infty} i^{2\tau}\theta_i^2 \leq \kappa \right\}.$$

Condition 4. *$\eta = \sum_{j\in S^\star} \eta_j$, and for all $j = 1,\dots,d$, $\eta_j \in \mathcal{W}(\tau,\kappa)$.*

In other words, we now assume that a sparse (additive) representation of η does exist with some sufficient regularity, and that its support is some ensemble $S^\star \subset \{1,\dots,d\}$.

We are now in a position to state our next result, which is again an adaptation of Theorem 2.

Theorem 5. *For any distribution of the random variables $(\mathbf{X}, Y)$, Condition 1 holds with $\psi = \delta^2/4n$. With the prior π defined as in (4), under Condition 2 and Condition 4, and choosing $\delta = \Upsilon n^{\frac{1+\tau}{1+2\tau}}$, we obtain for any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \left\{ C_1 n^{-\frac{\tau}{2\tau+1}} + C_2\left(2\log(2/\varepsilon) + K + |S^\star|_0 \log(C_3 n^{\frac{\tau+1}{2\tau+1}})\right) n^{-\frac{\tau+1}{2\tau+1}} \right\}\right] \geq 1-\varepsilon,$$

where Υ, C_1, C_2 and C_3 are constants depending only on c, C, β, κ, $|S^\star|_0$ and τ, and

$$K = 2\left(|S^\star|_0 \log\frac{de}{|S^\star|_0} + \log\frac{1}{1-\beta}\right).$$

The leading term $\mathcal{O}\left(n^{-\frac{\tau}{2\tau+1}}\right)$ is the classical nonparametric rate of convergence for the estimation of a function with some regularity τ, and the other terms involve the dimension of the best approaching model: in that sense, if a sparse representation of η exists, these terms will be small. Note that this result proves that our estimator $\hat{s}$ is adaptive to the unknown regularity τ and to the sparsity pattern $S^\star$.

Our last oracle inequality is our most detailed result, and combines the settings of Theorem 4 and Theorem 5.

Theorem 6. *For any distribution of the random variables $(\mathbf{X}, Y)$ satisfying Condition 3 for some parameter $\alpha \in (0,1)$, Condition 1 holds with $\psi = \frac{n}{2}\mathbb{V}(T(s))\phi\left(\frac{2\delta}{n}\right)$. Hence, using Theorem 2, under Condition 2 and Condition 4 and choosing $\delta = \Upsilon n^{\frac{1+\tau(1+\alpha)}{1+\tau(2+\alpha)}}$, we have for any $\varepsilon \in (0,1)$,*

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \left\{C_1 n^{\frac{-\tau(1+\alpha)}{1+(2+\alpha)\tau}} + C_2\left(2\log(2/\varepsilon) + K + |S^\star|_0 \log\left(C_3 n^{\frac{1+\tau(1+\alpha)}{1+\tau(2+\alpha)}}\right)\right) n^{\frac{-(1+\tau(1+\alpha))}{1+(2+\alpha)\tau}}\right\}\right] \geq 1-\varepsilon,$$

where Υ, C_1, C_2 and C_3 are constants depending only on c, C, β, κ, $|S^\star|_0$, τ, $\mathcal{C}$ and α, and

$$K = 2\left(|S^\star|_0 \log \frac{de}{|S^\star|_0} + \log \frac{1}{1-\beta}\right).$$

Again, note that this result proves our estimator $\hat{s}$ to be fully adaptive to the unknown smoothness τ, to the margin parameter α *and* to the unknown sparsity pattern $S^\star$. Let us conclude this section by a comment on the links between our results and the minimax results introduced in Clémençon and Robbiano (2011). To the best of our knowledge, Clémençon and Robbiano (2011) are the first to prove minimax optimality results for the bipartite ranking problem, with oracle inequalities in expectation. The minimax rate of convergence is exactly the one appearing in Theorem 6, namely

$$\mathcal{O}\left(n^{\frac{-(1+\tau(1+\alpha))}{1+(2+\alpha)\tau}}\right).$$

Since our results hold in probability, it is straightforward to obtain the similar oracle inequalities in expectation (integrating with respect to ε). Our PAC-Bayesian estimator thus achieves the minimax rate of convergence for the bipartite ranking problem, moreover on a larger functional class (Sobolev ellipsoid vs. Hölder class).

4 MCMC implementation

Our approach requires to sample from the Gibbs posterior $\hat{\rho}_\delta$ and we rely on an MCMC procedure to do so. However, several pitfalls appear: since

$\hat{\rho}_\delta$ is a distribution on a very high dimensional space with a complex structure, classical MCMC algorithms are likely to perform poorly. We therefore propose an adaptation to the ranking setting of the algorithm presented in Guedj and Alquier (2013) in the context of regression, which is inspired by Petralias and Dellaportas (2012). The key idea lies in the definition of a neighborhood relationship among the different models. Let us recall that $\mathbf{m} = (m_1, \dots, m_d) \in \{0,1\}^d$ denotes a model. Our transdimensional gateway is defined as follows: at each MCMC step, we propose to add a missing covariate to the current model, delete an existing one, or keep the same covariates. This entices the definition of three possible neighborhoods for a model $\mathbf{m}^t$:

- The set of all models having the same covariates that $\mathbf{m}^t$, plus one, denoted $\mathcal{V}^+_{\mathbf{m}_t}$,
- The set of all models having the same covariates that $\mathbf{m}^t$, minus one, denoted $\mathcal{V}^-_{\mathbf{m}_t}$,
- The neighborhood corresponding to the case where no dimension change is proposed at iteration t, which is limited to the current model $\mathbf{m}^t$.

We first select a neighborhood (*i.e.*, a move) among $\mathcal{V}^+_{\mathbf{m}_t}$, $\mathcal{V}^-_{\mathbf{m}_t}$ and $\{\mathbf{m}^t\}$ with probabilities (a, a, b) (*e.g.*, $a = 1/4$ and $b = 1/2$ in the following). Let $\mathcal{V}$ denote the selected neighborhood. For any model in $\mathcal{V}$, a candidate vector θ is sampled from a proposal Gaussian distribution, whose mean is a benchmark estimator (such as least-squares fit, the maximum likelihood estimator, a Lasso estimator, etc.) and whose variance is a parameter to the algorithm. The joint move towards the candidate model and estimator is then accepted following a Metropolis-Hastings ratio. This approach has been implemented for the regression problem in Guedj (2013).

Note that this algorithm is designed to sample from $\hat{\rho}_\delta$. However, estimators sampled from this algorithm may suffer from large variance, and we introduce a more stable version of our algorithm. Recall that in (2), $\hat{\theta}$ is sampled from $\hat{\rho}_\delta$. From a numerical stability perspective, it is useful to consider the mean of $\hat{\rho}_\delta$ instead. Thus we adapt our notation and now define two different estimators: $s_{\hat{\theta}^r}$ (same estimator as in (2)) and $s_{\hat{\theta}^a}$, where

$$\hat{\theta}^r \sim \hat{\rho}_\delta \quad \text{(randomized estimator)}, \tag{5}$$

and

$$\hat{\theta}^a = \int_\Theta \theta \hat{\rho}_\delta(\mathrm{d}\theta) = \mathbb{E}_{\hat{\rho}_\delta}\theta \quad \text{(averaged estimator)}. \tag{6}$$

In the following results, estimators defined in (5) and (6) will be referred to as *PAC-Bayesian Randomized* and *PAC-Bayesian Averaged*, respectively. Finally, we introduce the following notation: for any model $\mathbf{m}$, we let $\theta_{\mathbf{m}}$ denote

a benchmark estimator, $\varphi_{\mathbf{m}}$ denotes the density of the Gaussian distribution $\mathcal{N}(\theta_{\mathbf{m}}, \sigma^2 I_{\mathbf{m}})$ where $I_{\mathbf{m}}$ stands for the identity matrix $|\mathbf{m}|_0 \times |\mathbf{m}|_0$.
The pseudocode is presented in Algorithm 1.

Algorithm 1 An MCMC algorithm for PAC-Bayesian estimators

Input:
horizon T,
burnin b,
proposal variance $\sigma^2 > 0$,
inverse temperature parameter $\delta > 0$.

Output: two sequences of models $(\mathbf{m}^t)_{t=1}^T$ and estimators $(\theta^t)_{t=1}^T$.

At time $t = 2, \ldots, T$,

1: Pick a move and form the corresponding neighborhood $\mathcal{V}_t$.

2: For all $\mathbf{m} \in \mathcal{V}_t$, draw a candidate estimator $\tilde{\theta}_{\mathbf{m}} \sim \mathcal{N}(\theta_{\mathbf{m}}, \sigma^2 I_{\mathbf{m}})$.

3: Pick a pair $(\mathbf{m}, \tilde{\theta}_{\mathbf{m}})$ with probability proportional to $\hat{\rho}_\delta(\tilde{\theta}_{\mathbf{m}}) / \varphi_{\mathbf{m}}(\tilde{\theta}_{\mathbf{m}})$.

4: Set

$$\begin{cases} \theta^t = \tilde{\theta}_{\mathbf{m}} \\ \mathbf{m}^t = \mathbf{m} \end{cases} \text{with probability } \alpha, \qquad \begin{cases} \theta^t = \theta^{t-1} \\ \mathbf{m}^t = \mathbf{m}^{t-1} \end{cases} \text{with probability } 1-\alpha,$$

$$\alpha = \min\left(1, \frac{\hat{\rho}_\delta(\tilde{\theta}_{\mathbf{m}}) \varphi_{\mathbf{m}}(\theta^{t-1})}{\hat{\rho}_\delta(\theta^{t-1}) \varphi_{\mathbf{m}}(\tilde{\theta}_{\mathbf{m}})}\right).$$

Final estimators:

$$s_{\hat{\theta}^r} : \mathbf{X} \mapsto \sum_{j=1}^d \sum_{k=1}^M \theta_{jk}^T \phi_k(X_j) \qquad \text{(PAC-Bayesian Randomized)},$$

$$s_{\hat{\theta}^a} : \mathbf{X} \mapsto \sum_{j=1}^d \sum_{k=1}^M \left(\sum_{\ell=b+1}^T \theta_{jk}^\ell \right) \phi_k(X_j) \qquad \text{(PAC-Bayesian Averaged)}.$$

Recall that our procedure relies on the dictionary $\mathbb{D}$. In our implementation, we chose $M = 13$ (a value achieving a compromise between computational feasibility and analytical flexibility) and as functions the seven first Legendre polynomials and the six first trigonometric polynomials.

The algorithm mainly depends on two input parameters, the variance of the proposal distributions σ^2 and the inverse temperature parameter $\delta > 0$. Bad choices for these two parameters are likely to quickly deteriorates the performance of the algorithm. Indeed, if σ^2 is large, the proposal gaussian distribution will generate candidate estimators weakly related to the benchmark ones. On the contrary, small values for σ^2 will concentrate candidates towards benchmark estimators, whittling the diversity of estimators proposed. It is important to note that proposing randomized candidate estimators is a key part of our work.

As for the inverse temperature parameter δ, small values clearly make the Gibbs posterior very similar to the sparsity-inducing prior. In that case, fit to the data is negligible, and the performance are likely to drop. On the contrary, large values for δ will concentrate most of the mass of the Gibbs posterior towards a minimizer of the empirical ranking risk, which is possibly non-sparse. For these reasons, we conducted a thorough study of the influence of the two parameters σ^2 and δ, on the following synthetic model.

$$n_{\text{train}} = 1000, \quad n_{\text{test}} = 2000, \quad d = 10, \quad \mathbf{X} \sim \mathcal{U}([0,1]^d),$$
$$Y = \mathbb{1}\{\eta(\mathbf{X}) > \mathcal{U}([0,1])\}, \quad \eta : \mathbf{x} \mapsto X_3 + 7X_3^2 + 8\sin(\pi X_5) \quad (7)$$

Note that the values taken by η have been renormalized to fit in $(0,1)$. We take as a benchmark in our simulations the AUC of η as defined in (7), which is .7387. In other words, the closer to .7387, the better the performance. Our simulations are summed up in Table **??** and Figure 2. Several comments are in order.

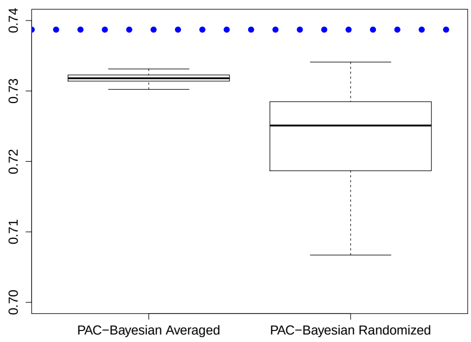
- The overall performance is good, and the best AUC value (equal to .731, achieved by the averaged estimator with $\delta = .1$ and $\sigma^2 = .001$) is very close to the optimal oracle value .7387, therefore validating our procedure for ranking.
- As expected, the averaged estimator exhibits slightly better performance (see Figure 2, a), due to its improved numerical stability over the randomized estimator.
- δ and σ^2 require fine tuning. Setting one of these two parameters at a wrong magnitude will fail. We advise to consider methods such as cross-validation to perform automatic calibration.
- When δ and σ^2 are finely calibrated, the algorithm almost always selects the two ground truth covariates (3 and 5, see Figure 2, b). As highlighted above, additional junk covariates may be selected when the fit to the data is weak (*i.e.*, δ is too small), leading to poor performance.

Table 1 – Mean (and variance) of AUC over 50 replications for both estimators (5) and (6) (1000 MCMC iterations, 800 burnin iterations), and frequencies of the selected variables for the last 200 iterations.

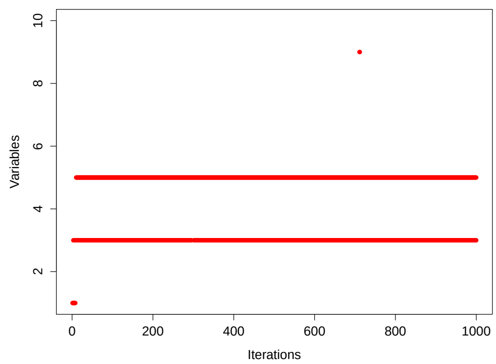
δ	σ^2	PAC-Bayesian Averaged	PAC-Bayesian Randomized	#3	#5	$\sum_{i \neq \{3,5\}} \#i$
100	1	.699 (.020)	.697 (.019)	.9363	.9200	.0272
100	.1	.713 (.014)	.712 (.014)	.9881	.9800	.0369
100	.01	.723 (.009)	.721 (.010)	1.0000	1.0000	.0617
100	.001	.720 (.006)	.712 (.006)	1.0000	1.0000	.4251
10	1	.701 (.024)	.702 (.024)	.9585	.9000	.0396
10	.1	.713 (.012)	.712 (.014)	1.0000	1.0000	.0561
10	.01	.724 (.008)	.724 (.008)	1.0000	1.0000	.0404
10	.001	.721 (.006)	.719 (.006)	1.0000	1.0000	.3970
1	1	.705 (.019)	.705 (.014)	1.0000	1.0000	.0898
1	.1	.715 (.012)	.714 (.012)	1.0000	1.0000	.0672
1	.01	.725 (.004)	.725 (.004)	1.0000	1.0000	.0286
1	.001	.727 (.003)	.725 (.004)	1.0000	1.0000	.1443
.1	1	.703 (.014)	.700 (.014)	1.0000	1.0000	.0977
.1	.1	.716 (.010)	.709 (.013)	.9996	.9981	.0304
.1	.01	.729 (.003)	.717 (.010)	.9998	.9982	.0069
.1	.001	.731 (.001)	.723 (.006)	.9995	.9996	.0004
.01	1	.663 (.045)	.563 (.058)	.5323	.4641	.1913
.01	.1	.667 (.058)	.561 (.068)	.3811	.4426	.0620
.01	.01	.672 (.017)	.634 (.028)	.4293	.5783	.0059
.01	.001	.663 (.019)	.646 (.042)	.5743	.3700	.0089

Figure 2 – Numerical experiments on synthetic data. The two meaningful covariates are the third and the fifth.

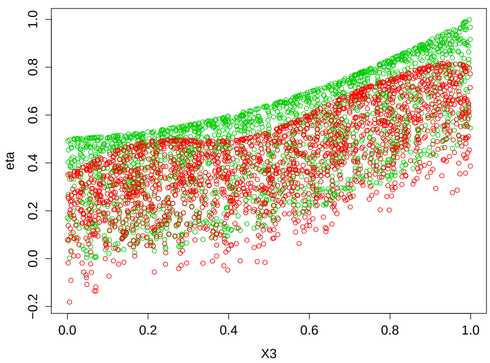
(a) Boxplot of AUC for both estimators. Blue dotted line: optimal oracle value (.7387).



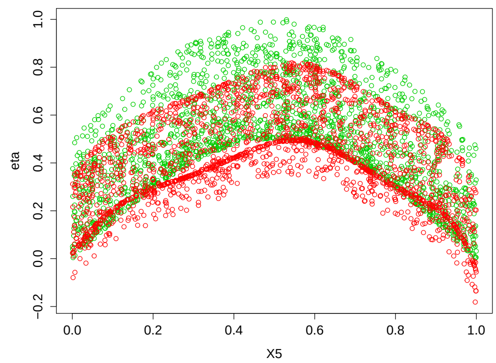
(b) Selected variables along the MCMC chain.



(c) For the third covariate, green points are $\eta(X_i)$ and red ones are $s_{\hat{\theta}^a}(X_i)$.



(d) For the fifth covariate, green points are $\eta(X_i)$ and red ones are $s_{\hat{\theta}^a}(X_i)$.



We now compare our PAC-Bayesian procedure to two state-of-the-art methods, on real-life datasets. Since our work investigates nonlinear scoring functions, we restricted the comparison with similar methods. The Rankboost (Freund et al., 2003) and TreeRank (Baskiotis et al., 2010; Clémençon et al., 2011) algorithms appear as ideal benchmarks. We have conducted a series of experiments on the following datasets: Diabetes, Heart, Iono, Messidor, Pima and Spectf. All these datasets are freely available online, following http://archive.ics.uci.edu/ml/ and serve as classical benchmark for machine learning tasks. Our results are wrapped up in Table 2 (where TRT, TRl and TRg denote the TreeRank algorithm trained with decision trees, linear SVM and gaussian SVM, respectively). For most datasets, our PAC-Bayesian estimators compete on similar ground with the four other methods, intercalating between the less and most performant method, while being the only ones supported by ground theoretical results.

5 Conclusion

We study in the present paper the problem of bipartite ranking in its theoretical and algorithmic aspects, in a high-dimensional setting through the PAC-Bayesian approach. Our model is nonlinear and assumes a sparse additive representation of the optimal scoring function. We propose an estimator based on the Gibbs pseudo-posterior distribution, and derive oracle inequalities in probability under the sparsity assumption, *i.e.*, with terms involving the intrinsic dimension instead of the ambient dimension d. Under minimal assumption, we recover classical rates of convergence $\mathcal{O}(n^{-1/2})$. On a Sobolev ellipsoid with regularity τ and under a margin assumption of parameter α, we obtain the minimax rate of convergence

$$\mathcal{O}\left(n^{\frac{-(1+\tau(1+\alpha))}{1+(2+\alpha)\tau}}\right).$$

A salient fact is that our results significantly extend previous works (Clémençon and Robbiano, 2011), in that our inequalities hold in probability (instead of inequalities in expectation). Next, we propose an implementation of the PAC-Bayesian estimators through a transdimensional MCMC. Its performance on synthetic and real-life datasets competes with other nonlinear ranking algorithms. In conclusion, the main contributions of this paper are a nonlinear procedure with provable minimax rates and an performant implemented approximation for the bipartite ranking problem.

Table 2 – Cross-validated mean (and variance) of AUC over seven real-life datasets.

	PAC-Bayesian Averaged	PAC-Bayesian Randomized	TRT	TRl	TRg	Rankboost
Vowel	.848 (.018)	.846 (.017)	.908 (.017)	.946 (.011)	.976 (.009)	.946 (.013)
Messidor	.738 (.033)	.732 (.034)	.687 (.038)	.800 (.023)	.754 (.036)	.747 (.021)
Iono	.871 (.047)	.869 (.047)	.878 (.033)	.846 (.050)	.905 (.024)	.929 (.017)
Diabetes	.772 (.038)	.767 (.040)	.777 (.037)	.810 (.036)	.794 (.037)	.820 (.033)
Heart	.733 (.061)	.725 (.065)	.700 (.072)	.752 (.063)	.676 (.077)	.746 (.062)
Pima	.782 (.036)	.772 (.038)	.777 (.037)	.810 (.036)	.703 (.033)	.820 (.033)
Spectf	.764 (.103)	.759 (.102)	.757 (.129)	.711 (.109)	.601 (.089)	.854 (.111)

6 Proofs

The following lemma (Legendre transform of the Kullback-Leibler divergence, Csiszár, 1975) is a key ingredient in our proofs and the demonstration may be found in Catoni (2004, Equation 5.2.1).

Lemma 2. *Let $(A, \mathcal{A})$ be a measurable space. For any probability μ on $(A, \mathcal{A})$ and any measurable function $h : A \to \mathbb{R}$ such that $\int (\exp \circ h) \mathrm{d}\mu < \infty$,*

$$\log \int (\exp \circ h) \mathrm{d}\mu = \sup_{m \in \mathcal{M}^1_{+,\pi}(A,\mathcal{A})} \left\{ \int h \mathrm{d}m - \mathcal{K}(m, \mu) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ, the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{\mathrm{d}g}{\mathrm{d}\mu}(a) = \frac{\exp \circ h(a)}{\int (\exp \circ h) \mathrm{d}\mu}, \quad a \in A.$$

Proof of Theorem 1. Under Condition 1, for any $\varepsilon \in (0,1)$, we have

$$\mathbb{E} \exp \left[\delta(\mathcal{E}_n(s) - \mathcal{E}(s)) - \psi(s) - \log(1/\varepsilon) \right] \leq \varepsilon.$$

Now, assume that s is drawn from some prior distribution π. We can integrate on both sides of the previous inequality, thus

$$\int \left\{ \mathbb{E} \exp \left[\delta(\mathcal{E}_n(s) - \mathcal{E}(s)) - \psi(s) - \log(1/\varepsilon) \right] \right\} \pi(\mathrm{d}s) \leq \varepsilon.$$

Using a Fubini-Tonelli theorem, we may write

$$\mathbb{E} \int \exp \left[\delta(\mathcal{E}_n(s) - \mathcal{E}(s)) - \psi(s) - \log(1/\varepsilon) \right] \pi(\mathrm{d}s) \leq \varepsilon.$$

Now, let ρ denote an absolutely continuous distribution with respect to π. We obtain

$$\mathbb{E} \int \exp \left[\delta(\mathcal{E}_n(s) - \mathcal{E}(s)) - \psi(s) - \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(s) - \log(1/\varepsilon) \right] \rho(\mathrm{d}s) \leq \varepsilon.$$

With a slight extension of previous notation, we let $\mathbb{E}$ stands for the expectation computed with respect to the distribution of $(\mathbf{X}, Y)$ *and* the posterior distribution ρ. Using the elementary inequality $\exp(\delta x) \geq \mathbb{1}_{\mathbb{R}_+}(x)$,

$$\mathbb{P} \left[\mathcal{E}_n(s) - \mathcal{E}(s) - \frac{\psi + \log(1/\varepsilon) + \log \frac{\mathrm{d}\rho}{\mathrm{d}\pi}(s)}{\delta} > 0 \right] \leq \varepsilon,$$

i.e.,

$$\mathbb{P}\left[\mathcal{E}_n(s) \leq \mathcal{E}(s) + \frac{\psi(s) + \log(1/\varepsilon) + \log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(s)}{\delta}\right] \geq 1-\varepsilon. \tag{8}$$

We may now apply the same scheme of proof with the variables $\widetilde{T}_{i,j} = -T_{i,j}$ to obtain

$$\mathbb{P}\left[\mathcal{E}(s) \leq \mathcal{E}_n(s) + \frac{\psi(s) + \log(1/\varepsilon) + \log\frac{\mathrm{d}\rho}{\mathrm{d}\pi}(s)}{\delta}\right] \geq 1-\varepsilon.$$

Now, for the choice $\rho = \hat{\rho}_\delta$ and $\hat{s} \sim \hat{\rho}_\delta$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \mathcal{E}_n(\hat{s}) + \frac{\psi(\hat{s}) + \log(1/\varepsilon) + \log\frac{\mathrm{d}\hat{\rho}_\delta}{\mathrm{d}\pi}(\hat{s})}{\delta}\right] \geq 1-\varepsilon.$$

Note that

$$\begin{aligned}\log\frac{\mathrm{d}\hat{\rho}_\delta}{\mathrm{d}\pi}(\hat{s}) &= \log\frac{\exp(-\delta L_n(\hat{s}))}{\int\exp(-\delta L_n(s'))\pi(\mathrm{d}s')}\\ &= -\delta L_n(\hat{s}) - \log\int\exp\left(-\delta L_n(s')\right)\pi(\mathrm{d}s').\end{aligned}$$

Hence

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq -L_n(\eta) + \frac{1}{\delta}\left(\psi(\hat{s}) + \log(1/\varepsilon) - \log\int\exp(-\delta L_n(s'))\pi(\mathrm{d}s')\right)\right] \geq 1-\varepsilon.$$

From Lemma 2, we obtain

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int\mathcal{E}_n(s)\rho(\mathrm{d}s) + \frac{\psi(\hat{s}) + \log(1/\varepsilon) + \mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon,$$

where $s \sim \rho$. So by integrating (8),

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int\mathcal{E}(s)\rho(\mathrm{d}s) + \int\frac{\psi(s)}{\delta}\rho(\mathrm{d}s) + \frac{\psi(\hat{s}) + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon, \tag{9}$$

which is the desired result. □

Proof of Corollary 1. For some candidate function s and any $i,j = 1,\ldots,n$, define

$$T_{i,j} = \mathbb{1}_{\{(s(\mathbf{X}_i)-s(\mathbf{X}_j))(Y_i-Y_j)<0\}} - \mathbb{1}_{\{(\eta(\mathbf{X}_i)-\eta(\mathbf{X}_j))(Y_i-Y_j)<0\}}.$$

Using results on U-statistics (Hoeffding decomposition of U-statistics, Serfling, 1980), we may write, for any $\gamma > 0$,

$$\begin{aligned}
&\mathbb{E}\exp\left[\gamma \sum_{i\neq j}(T_{i,j} - \mathbb{E}T_{i,j})\right] \\
&\quad = \mathbb{E}\exp\left[\frac{\gamma n(n-1)}{n!}\sum_{\pi}\frac{1}{n/2}\sum_{i=1}^{n/2}(T_{\pi(i),\pi(i+n/2)} - \mathbb{E}T_{\pi(i),\pi(i+n/2)})\right] \\
&\quad \leq \mathbb{E}\exp\left[2\gamma(n-1)\sum_{i=1}^{n/2}(T_{i,i+n/2} - \mathbb{E}T_{i,i+n/2})\right], \qquad (10)
\end{aligned}$$

where we used the Jensen's inequality. Next, using an independence argument and Hoeffding's inequality applied to the random variable $T_{i,i+n/2} - \mathbb{E}T_{i,i+n/2} \in (-\mathbb{E}T_{i,i+n/2}, 1 - \mathbb{E}T_{i,i+n/2})$,

$$\begin{aligned}
\mathbb{E}\exp\left[\gamma \sum_{i\neq j}(T_{i,j} - \mathbb{E}T_{i,j})\right] &= \prod_{i=1}^{n/2}\mathbb{E}\exp\left[2\gamma(n-1)(T_{i,i+n/2} - \mathbb{E}T_{i,i+n/2})\right] \\
&\leq \prod_{i=1}^{n/2}\exp\left(\frac{\gamma^2(n-1)^2}{2}\right) \\
&= \exp\left(\frac{n\gamma^2(n-1)^2}{4}\right) \\
&= \exp(\psi),
\end{aligned}$$

with $\psi = \delta^2/4n$ and $\delta = \gamma n(n-1)$. Note that ψ does not depend on s. Finally, note that

$$\mathbb{E}\exp\left[\gamma \sum_{i\neq j}(T_{i,j} - \mathbb{E}T_{i,j})\right] = \mathbb{E}\exp[\delta(\mathcal{E}_n(s) - \mathcal{E}(s))].$$

Corollary 1 is then a straightforward application of Theorem 1. □

Proof of Theorem 2. From Theorem 1 and since $\psi = \delta^2/4n$ does not depend on s, considering $\rho_{\mathbf{m}}$ as a probability measure whose support is $\mathbb{R}^{|\mathbf{m}|_0}$,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}}\inf_{\rho_{\mathbf{m}}}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \frac{2\psi + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho_{\mathbf{m}},\pi)}{\delta}\right\}\right] \geq 1 - \varepsilon.$$

Next, note that

$$\begin{aligned}
\mathcal{K}(\rho_{\mathbf{m}},\pi) &= \mathcal{K}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + |\mathbf{m}|_0 M\log(1/\beta) + \log\binom{d}{|\mathbf{m}|_0} + \log\frac{1-\beta^{d+1}}{1-\beta} \\
&\leq \mathcal{K}(\rho_{\mathbf{m}},\pi_{\mathbf{m}}) + |\mathbf{m}|_0 M\log(1/\beta) + |\mathbf{m}|_0\log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}, \qquad (11)
\end{aligned}$$

where we used the elementary inequality $\log\binom{d}{k} \le k \log \frac{de}{k}$. Note that if we consider as distributions $\rho_{\mathbf{m}}$ uniform distributions on ℓ_2-balls centered in any $\theta \in \mathcal{B}_{\mathbf{m}}$ such that $\|\theta\| = 1$, of radius $t \in (0,1)$, we obtain that

$$\mathcal{K}(\rho_{\mathbf{m}}, \pi_{\mathbf{m}}) = |\mathbf{m}|_0 \log(1/t).$$

For some θ_0 such that $\|\theta_0\| = 1$,

$$\begin{aligned}
R(s_\theta) &= \mathbb{E}[\mathbb{1}_{\{(s_\theta(\mathbf{X})-s_\theta(\mathbf{X}'))(Y-Y')<0\}}] \\
&= \mathbb{E}[\mathbb{1}_{\{(s_{\theta_0}(\mathbf{X})-s_{\theta_0}(\mathbf{X}'))(Y-Y')<0\}}] \\
&\qquad + \mathbb{E}[\mathbb{1}_{\{(s_\theta(\mathbf{X})-s_\theta(\mathbf{X}'))(Y-Y')<0\}} - \mathbb{1}_{\{(s_{\theta_0}(\mathbf{X})-s_{\theta_0}(\mathbf{X}'))(Y-Y')<0\}}] \\
&\le R(s_{\theta_0}) + \mathbb{P}[\operatorname{sign}(s_\theta(\mathbf{X}) - s_\theta(\mathbf{X}'))(Y-Y')) \neq \operatorname{sign}(s_{\theta_0}(\mathbf{X}) - s_{\theta_0}(\mathbf{X}'))(Y-Y')] \\
&= R(s_{\theta_0}) + \mathbb{P}[\operatorname{sign}(s_\theta(\mathbf{X}) - s_\theta(\mathbf{X}')) \neq \operatorname{sign}(s_{\theta_0}(\mathbf{X}) - s_{\theta_0}(\mathbf{X}'))] \\
&\le R(s_{\theta_0}) + 2c\|\theta - \theta_0\|,
\end{aligned}$$

where we used Condition 2 in the last inequality. Let $\rho_{\mathbf{m},\theta_0,t}$ denote the uniform distribution on the ℓ_2-ball centered in θ_0 and of radius $t \in (0,1)$. From what precedes,

$$\int \mathcal{E}(s_\theta)\rho_{\mathbf{m},\theta_0,t}(\mathrm{d}s) = R(s_{\theta_0}) + 2ct. \tag{12}$$

Using the notation

$$K = 2\left(|\mathbf{m}|_0 M \log(1/\beta) + |\mathbf{m}|_0 \log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}\right), \tag{13}$$

we obtain

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \inf_{t\in(0,1)} \left\{R(s_{\theta_0}) + 2ct \right.\right.$$
$$\left.\left. + \frac{2\psi + 2\log(2/\varepsilon) + 2|\mathbf{m}|_0\log(1/t) + 2K}{\delta}\right\}\right] \ge 1-\varepsilon.$$

It can easily be seen that the function $t \mapsto 2ct + \log(1/t)/\delta$ is upper-bounded at the point $t = 1/(2c\delta)$. Therefore,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \left\{R(s_{\theta_0}) + \frac{1 + 2\psi + 2\log(2/\varepsilon) + 2|\mathbf{m}|_0\log(2c\delta) + 2K}{\delta}\right\}\right]$$
$$\ge 1-\varepsilon.$$

Now, recalling that $\psi = \delta^2/4n$, we choose $\delta = \sqrt{n}$. The desired result is then straightforward from the proof of Theorem 1. □

Proof of Lemma 1. First, note that

$$\mathbb{E}T = L(s) - L^\star \geq 0.$$

Let

$$\begin{aligned}
r &= r(\mathbf{X},\mathbf{X}') = \operatorname{sign}(s(\mathbf{X}) - s(\mathbf{X}')),\\
r^\star &= r^\star(\mathbf{X},\mathbf{X}') = \operatorname{sign}(\eta(\mathbf{X}) - \eta(\mathbf{X}')),\\
Z &= (Y - Y')/2.
\end{aligned}$$

With this notation,

$$\begin{aligned}
\mathbb{E}T^2 &= \mathbb{E}\big[\mathbb{1}_{\{r\neq Z\}} + \mathbb{1}_{\{r^\star\neq Z\}} - 2\mathbb{1}_{\{r\neq Z\}}\mathbb{1}_{\{r^\star\neq Z\}}\big]\\
&= \mathbb{E}\big[\mathbb{1}_{\{r=1\}}\mathbb{1}_{\{Z=-1\}} + \mathbb{1}_{\{r=-1\}}\mathbb{1}_{\{Z=1\}} + \mathbb{1}_{\{r^\star=1\}}\mathbb{1}_{\{Z=-1\}} + \mathbb{1}_{\{r^\star=-1\}}\mathbb{1}_{\{Z=1\}}\\
&\qquad - 2\mathbb{1}_{\{r=1\}}\mathbb{1}_{\{r^\star=1\}}\mathbb{1}_{\{Z=-1\}} - 2\mathbb{1}_{\{r=-1\}}\mathbb{1}_{\{r^\star=-1\}}\mathbb{1}_{\{Z=1\}}\big].
\end{aligned}$$

Next,

$$\begin{aligned}
\mathbb{E}T^2 = \mathbb{E}\big[&\mathbb{1}_{\{r=1\}}(1-\eta(\mathbf{X}))\eta(\mathbf{X}') + \mathbb{1}_{\{r=-1\}}\eta(\mathbf{X})(1-\eta(\mathbf{X}')) + \mathbb{1}_{\{r^\star=1\}}(1-\eta(\mathbf{X}))\eta(\mathbf{X}')\\
&+ \mathbb{1}_{\{r^\star=-1\}}\eta(\mathbf{X})(1-\eta(\mathbf{X}')) - 2\mathbb{1}_{\{r=1\}}\mathbb{1}_{\{r^\star=1\}}(1-\eta(\mathbf{X}))\eta(\mathbf{X}')\\
&- 2\mathbb{1}_{\{r=-1\}}\mathbb{1}_{\{r^\star=-1\}}\eta(\mathbf{X})(1-\eta(\mathbf{X}'))\big].
\end{aligned}$$

Thus,

$$\begin{aligned}
\mathbb{E}T^2 &= \mathbb{E}\big[(1-\eta(\mathbf{X}))\eta(\mathbf{X}')(\mathbb{1}_{\{r=1\}} + \mathbb{1}_{\{r^\star=1\}} - 2\mathbb{1}_{\{r=1\}}\mathbb{1}_{\{r^\star=1\}})\\
&\qquad + \eta(\mathbf{X})(1-\eta(\mathbf{X}'))(\mathbb{1}_{\{r=-1\}} + \mathbb{1}_{\{r^\star=1\}} - 2\mathbb{1}_{\{r=-1\}}\mathbb{1}_{\{r^\star=-1\}})\big]\\
&= \mathbb{E}\big[\mathbb{1}_{\{r\neq r^\star\}}((1-\eta(\mathbf{X}))\eta(\mathbf{X}') + \eta(\mathbf{X})(1-\eta(\mathbf{X}')))\big]\\
&= \mathbb{E}\big[\mathbb{1}_{\{r\neq r^\star\}}(\eta(\mathbf{X}) + \eta(\mathbf{X}') - 2\eta(\mathbf{X})\eta(\mathbf{X}'))\big]\\
&\leq \frac{1}{2}\mathbb{E}\big[\mathbb{1}_{\{r\neq r^\star\}}\big]\\
&\leq \frac{1}{2}C(L(s) - L^\star)^{\alpha/(1+\alpha)}.
\end{aligned}$$

Finally,

$$\mathbb{V}(T) = \mathbb{E}T^2 - (\mathbb{E}T)^2 \leq \mathcal{C}(L(s) - L^\star)^{\alpha/(1+\alpha)},$$

where $\mathcal{C}$ is a constant, which completes the proof. □

Proof of Theorem 3. Recalling (10), with the notation $\phi(u) = e^u - u - 1$ and Benett's inequality, we get

$$\begin{aligned}
\mathbb{E}\exp\left[\gamma\sum_{i\neq j}\big(T_{i,j}(s) - \mathbb{E}T_{i,j}(s)\big)\right] &\leq \prod_{i=1}^{n/2}\exp\big(\mathbb{V}\big(T_{i,i+n/2}(s)\big)\phi\big(2(n-1)\gamma\big)\big)\\
&= \exp\left(\frac{n}{2}\mathbb{V}(T(s))\phi(2(n-1)\gamma)\right),
\end{aligned}$$

with $\gamma = \frac{\delta}{n(n-1)}$, which yields

$$\mathbb{E}\exp[\delta(\mathcal{E}_n(s) - \mathcal{E}(s))] \leq \exp(\psi),$$

with

$$\psi = \frac{n}{2}\mathbb{V}(T(s))\phi\left(\frac{2\delta}{n}\right),$$

achieving the first statement of Theorem 3. Now, combining this result and (9),

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \int \frac{n\mathbb{V}(T(s))\phi(2\delta/n)}{2\delta}\rho(\mathrm{d}s) \right.\right.$$
$$\left.\left. + \frac{n\mathbb{V}(T(\hat{s}))\phi(2\delta/n)/2 + 2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon.$$

Using the elementary inequality $\phi(x)/x \leq x$ for any $x \in (0,1)$ yields

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \int \frac{2\delta\mathbb{V}(T(s))}{n}\rho(\mathrm{d}s) + \frac{2\delta\mathbb{V}(T(\hat{s}))}{n} \right.\right.$$
$$\left.\left. + \frac{2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon.$$

Now, using Lemma 1,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int \mathcal{E}(s)\rho(\mathrm{d}s) + \int \frac{2\delta\mathcal{C}\mathcal{E}(s)^{\frac{\alpha}{1+\alpha}}}{n}\rho(\mathrm{d}s) + \frac{2\delta\mathcal{C}\mathcal{E}(\hat{s})^{\frac{\alpha}{1+\alpha}}}{n} \right.\right.$$
$$\left.\left. + \frac{2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon.$$

Thus, for any $x \geq 0$,

$$\mathbb{P}\left[\left(1 - \frac{2\delta\mathcal{C}x}{n}\right)\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int\left(1 + \frac{2\delta\mathcal{C}x}{n}\right)\mathcal{E}(s)\rho(\mathrm{d}s) \right.\right.$$
$$\left.\left. + \int \frac{2\delta\mathcal{C}}{n}\left(\mathcal{E}(s)^{\frac{\alpha}{1+\alpha}} - x\mathcal{E}(s)\right)\rho(\mathrm{d}s) + \frac{2\delta\mathcal{C}}{n}\left(\mathcal{E}(\hat{s})^{\frac{\alpha}{1+\alpha}} - x\mathcal{E}(\hat{s})\right) + \frac{2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right]$$
$$\geq 1-\varepsilon.$$

Clearly, the function $t \mapsto t^{\frac{\alpha}{1+\alpha}} - xt$ is upper bounded by $x^{-\alpha}\frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^\alpha$, so

$$\mathbb{P}\left[\left(1 - \frac{2\delta\mathcal{C}x}{n}\right)\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{\int\left(1 + \frac{2\delta\mathcal{C}x}{n}\right)\mathcal{E}(s)\rho(\mathrm{d}s) \right.\right.$$
$$\left.\left. + \int \frac{2\delta\mathcal{C}\mathcal{C}_\alpha}{n}x^{-\alpha}\rho(\mathrm{d}s) + \frac{2\delta\mathcal{C}\mathcal{C}_\alpha}{n}x^{-\alpha} + \frac{2\log(2/\varepsilon) + 2\mathcal{K}(\rho,\pi)}{\delta}\right\}\right] \geq 1-\varepsilon, \quad (14)$$

where $\mathcal{C}_\alpha = \frac{1}{1+\alpha}\left(\frac{\alpha}{1+\alpha}\right)^\alpha$. Next, we choose $x = \frac{n}{4\delta\mathcal{C}}$. The function

$$\delta \mapsto \mathcal{C}_\alpha \left(\frac{4\mathcal{C}}{n}\right)^{1+\alpha} \delta^{1+\alpha} + \frac{4}{\delta}$$

is upper-bounded at the point $\delta = \Upsilon n^{\frac{1+\alpha}{2+\alpha}}$ where $\Upsilon = 2((1+\alpha)\mathcal{C}_\alpha)^{-\frac{1}{2+\alpha}}(2\mathcal{C})^{-\frac{1+\alpha}{2+\alpha}}$. Thus, we obtain

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\rho\in\mathcal{M}_\pi}\left\{3\int \mathcal{E}(s)\rho(\mathrm{d}s) + n^{-\frac{1+\alpha}{2+\alpha}}\left[\mathcal{C}_\alpha(\mathcal{C}\Upsilon)^{1+\alpha} + \Upsilon^{-1}\left(\log(2/\varepsilon) + \mathcal{K}(\rho,\pi)\right)\right]\right\}\right] \geq 1-\varepsilon,$$

which concludes the proof.

□

Proof of Theorem 4. Note that from (14), choosing $x = \frac{n}{4\delta\mathcal{C}}$, we get

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}}\inf_{\rho_{\mathbf{m}}}\left\{3\int \mathcal{E}(s)\rho(\mathrm{d}s) + 2\mathcal{C}_\alpha\left(\frac{4\delta\mathcal{C}}{n}\right)^{1+\alpha} + \frac{4\log(2/\varepsilon) + 4\mathcal{K}(\rho_{\mathbf{m}},\pi)}{\delta}\right\}\right] \geq 1-\varepsilon.$$

Now, recall the result from (11).

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}}\inf_{\theta\in\mathcal{B}_{\mathbf{m}},\|\theta\|=1}\inf_{t\in(0,1)}\left\{3\int \mathcal{E}(s_\theta)\rho_{\mathbf{m},\theta,t}(\mathrm{d}s) + 2\mathcal{C}_\alpha\left(\frac{4\delta\mathcal{C}}{n}\right)^{1+\alpha} + \frac{4\log(2/\varepsilon)}{\delta} + 4\frac{|\mathbf{m}|_0\log(1/t) + |\mathbf{m}|_0 M\log(1/\beta) + |\mathbf{m}|_0\log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}}{\delta}\right\}\right] \geq 1-\varepsilon.$$

With the notation stated in (13) and the result (12),

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}}\inf_{\theta\in\mathcal{B}_{\mathbf{m}},\|\theta\|=1}\inf_{t\in(0,1)}\left\{3\mathcal{E}(s_\theta) + 4ct + 2\mathcal{C}_\alpha\left(\frac{4\delta\mathcal{C}}{n}\right)^{1+\alpha} + \frac{4\log(2/\varepsilon)}{\delta} + 4\frac{|\mathbf{m}|_0\log(1/t) + K}{\delta}\right\}\right] \geq 1-\varepsilon.$$

Clearly, the function $t \mapsto 4ct + \log(1/t)/\delta$ is upper-bounded at the point $t = 1/(4c\delta)$, so

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \inf_{\mathbf{m}}\inf_{\theta\in\mathcal{B}_{\mathbf{m}},\|\theta\|=1}\left\{3\mathcal{E}(s_\theta) + 2\mathcal{C}_\alpha\left(\frac{4\delta\mathcal{C}}{n}\right)^{1+\alpha} + \frac{4\log(2/\varepsilon)}{\delta} + 4\frac{|\mathbf{m}|_0\log(4c\delta) + K}{\delta}\right\}\right] \geq 1-\varepsilon. \quad (15)$$

Finally, noticing that the function

$$\delta \mapsto \mathcal{C}_\alpha \left(\frac{4\delta\mathcal{C}}{n}\right)^{1+\alpha} + \frac{1}{\delta}$$

is upper-bounded at the point $\delta = n^{\frac{1+\alpha}{2+\alpha}} \left(\frac{1}{4\mathcal{C}}\right)^{\frac{1+\alpha}{2+\alpha}} \left(\frac{1}{\mathcal{C}_\alpha(1+\alpha)}\right)^{\frac{1}{2+\alpha}}$, we obtain

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \left\{3\mathcal{E}(s_\theta) \right.\right.$$

$$\left.\left. + n^{-\frac{1+\alpha}{2+\alpha}}(4K + 4\log(2/\varepsilon) + \log(4cn^{\frac{1+\alpha}{2+\alpha}}))\mathcal{C}_\alpha^{\frac{1}{2+\alpha}} \left(\frac{1}{4\mathcal{C}}\right)^{-\frac{1+\alpha}{2+\alpha}} \left[(1+\alpha)^{-\frac{1+\alpha}{2+\alpha}} + (1+\alpha)^{\frac{1}{2+\alpha}}\right]\right\}\right]$$

$$\ge 1-\varepsilon,$$

which concludes the proof.

□

Proof of Theorem 5. We denote by η^{proj} the projection of η onto $\mathcal{S}_\Theta$. Let $\Gamma_{\eta^{\text{proj}}} = \left\{(\mathbf{x},\mathbf{x}')| \left(\eta^{\text{proj}}(\mathbf{x}) - \eta^{\text{proj}}(\mathbf{x}')\right)\left(\eta(\mathbf{x}) - \eta(\mathbf{x}')\right) < 0\right\}$. For any $\mathbf{X},\mathbf{X}' \in \Gamma_{\eta^{\text{proj}}}$, we have that $|\eta(\mathbf{X}) - \eta(\mathbf{X}')| \le |\eta^{\text{proj}}(\mathbf{X}) - \eta(\mathbf{X})| + |\eta^{\text{proj}}(\mathbf{X}') - \eta(\mathbf{X}')|$. Since (see Clémençon et al., 2008)

$$L\left(\eta^{\text{proj}}\right) - L^\star = \mathbb{E}\left[|\eta(\mathbf{X}) - \eta(\mathbf{X}')|\mathbb{1}\{(\mathbf{X},\mathbf{X}') \in \Gamma_{\eta^{\text{proj}}}\}\right],$$

we obtain $L(\eta^{\text{proj}}) - L^\star \le 2\mathbb{E}[|\eta^{\text{proj}}(\mathbf{X}) - \eta(\mathbf{X})|]$. Next, using the statement (11) combined with Theorem 1, result (12) and the choice $\psi = \delta^2/4n$ (which holds whatever the distribution of $(\mathbf{X},Y)$ may be), we get that

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \inf_{t\in(0,1)} \left\{\mathcal{E}(s_\theta) + 2ct + \frac{\delta}{2n} + \frac{2\log(2/\varepsilon) + |\mathbf{m}|_0\log(1/t) + 2K}{\delta}\right\}\right]$$

$$\ge 1-\varepsilon,$$

where

$$K = |\mathbf{m}|_0 M\log(1/\beta) + |\mathbf{m}|_0 \log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}.$$

Since the function $t \mapsto 2ct + \log(1/t)/\delta$ is upper-bounded at the point $t = 1/(2c\delta)$, hence

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \left\{\mathcal{E}(s_\theta) + \frac{\delta}{2n} + \frac{1 + 2\log(2/\varepsilon) + |\mathbf{m}|_0\log(2c\delta) + 2K}{\delta}\right\}\right]$$

$$\ge 1-\varepsilon,$$

Now,

$$
\begin{aligned}
\inf_{\theta\in\mathcal{B}_{\mathbf{m}},\|\theta\|=1} \mathcal{E}(s_\theta) &= \mathcal{E}(\eta^{\mathrm{proj}}) \\
&\le 2\mathbb{E}|\eta^{\mathrm{proj}}(\mathbf{X}) - \eta(\mathbf{X})| \\
&\le 2\sqrt{\mathbb{E}|\eta^{\mathrm{proj}}(\mathbf{X}) - \eta(\mathbf{X})|^2} \\
&\le 2\sqrt{\int \sum_{j\in S^\star}\sum_{k\ge M+1} |\theta^\star_{jk}|^2 \phi_k^2(x)\mu(\mathrm{d}x)} \\
&\le 2\sqrt{B}\sqrt{\sum_{j\in S^\star}\sum_{k\ge M+1} |\theta^\star_{jk}|^2},
\end{aligned}
$$

since the ϕ_k's are such that $\int|\phi_k(x)|\mu(\mathrm{d}x) \le B$ where $B>0$ is a numerical constant. Using the definition of the Sobolev ellipsoid, we obtain

$$
\inf_{\theta\in\mathcal{B}_{\mathbf{m}},\|\theta\|=1} \mathcal{E}(s_\theta) \le 2\sqrt{B|S^\star|_0\kappa}(1+M)^{-\tau}.
$$

Hence

$$
\begin{aligned}
\mathbb{P}\Bigg[\mathcal{E}(\hat{s}) \le \inf_{\mathbf{m}} \Bigg\{ 2\sqrt{B\kappa|S^\star|_0}(1+M)^{-\tau} + \frac{\delta}{2n} + \frac{|m|_0 M\log(1/\beta)}{\delta} \\
+ \frac{1+|\mathbf{m}|_0\log(2c\delta) + 2\log(2/\varepsilon) + 2K'}{\delta}\Bigg\}\Bigg] \ge 1-\varepsilon,
\end{aligned}
$$

where

$$
K' = |\mathbf{m}|_0 \log\frac{de}{|\mathbf{m}|_0} + \log\frac{1}{1-\beta}.
$$

Next, observe that we get rid of the remaining infimum by substituting $|S^\star|$ to $|\mathbf{m}|_0$. The function $t\mapsto 2\sqrt{B\kappa|S^\star|_0}(1+t)^{-\tau} + |S^\star|_0\log(1/\beta)t/\delta$ is upper-bounded at the point $t = \left(\frac{\sqrt{|S^\star|_0}\log(1/\beta)}{2\sqrt{B\kappa}\tau\delta}\right)^{-\frac{1}{1+\tau}} - 1$, which yields

$$
\mathbb{P}\left[\mathcal{E}(\hat{s}) \le \left\{\zeta\delta^{\frac{-\tau}{1+\tau}} + \frac{\delta}{2n} + \frac{1+2\log(2/\varepsilon)+2K'| + |S^\star|_0\log(2c\delta)}{\delta}\right\}\right] \ge 1-\varepsilon,
$$

where $\zeta = \left(\left(2\sqrt{B\kappa}\right)^{\frac{1}{1+\tau}} |S^\star|_0^{\frac{1+2\tau}{2+2\tau}} \log(1/\beta)^{\frac{\tau}{1+\tau}} \left(\tau^{-\frac{\tau}{1+\tau}} + \tau^{\frac{1}{1+\tau}}\right)\right)$, and

$$
K' = |S^\star|_0 \log\frac{de}{|S^\star|_0} + \log\frac{1}{1-\beta}.
$$

The function $t\mapsto \zeta t^{-\frac{\tau}{1+\tau}} + \frac{t}{2n}$ is upper-bounded at the point $t = n^{\frac{\tau+1}{2\tau+1}}\left(\frac{2\zeta\tau}{1+\tau}\right)^{\frac{\tau+1}{2\tau+1}}$, hence we obtain

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \left\{\zeta\left(\frac{2\zeta\tau n}{1+\tau}\right)^{\frac{\tau+1}{2\tau+1}\times\frac{-\tau}{1+\tau}} + \frac{1}{2n}\left(\frac{2\zeta\tau n}{1+\tau}\right)^{\frac{\tau+1}{2\tau+1}} \right.\right.$$
$$\left.\left. + \left(2\log(2/\varepsilon) + 2K' + |S^\star|_0 \log(2c\delta)\right)\left(\frac{2\zeta\tau n}{1+\tau}\right)^{-\frac{\tau+1}{2\tau+1}}\right\}\right] \geq 1-\varepsilon.$$

Finally,

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \leq \left\{\xi n^{-\frac{\tau}{2\tau+1}} + (2\log(2/\varepsilon) + 2K' + |S^\star|_0 \log(C_1 n^{\frac{\tau+1}{2\tau+1}}))\left(\frac{2\zeta\tau n}{1+\tau}\right)^{-\frac{\tau+1}{2\tau+1}}\right\}\right]$$
$$\geq 1-\varepsilon,$$

where $\xi = \zeta^{\frac{\tau+1}{2\tau+1}} 2^{-\frac{\tau}{2\tau+1}} \left(\left(\frac{\tau}{1+\tau}\right)^{-\frac{\tau}{2\tau+1}} + \left(\frac{\tau}{1+\tau}\right)^{\frac{\tau+1}{2\tau+1}}\right)$ and $C_1 = 2c\left(\frac{2\zeta\tau}{1+\tau}\right)^{\frac{\tau+1}{2\tau+1}}$.

□

Proof of Theorem 6. Observe that (as written in Clémençon and Robbiano, 2011, Proposition 4)

$$\begin{aligned}
\inf_{\theta\in\mathcal{B}_{\mathbf{m}}, \|\theta\|=1} \mathcal{E}(s_\theta) &= \mathcal{E}(\eta^{\text{proj}}) \\
&\leq 2\|\eta^{\text{proj}} - \eta\|_\infty \mathbb{P}\left[\left(\eta^{\text{proj}}(\mathbf{X}) - \eta^{\text{proj}}(\mathbf{X}')\right)\left(\eta(\mathbf{X}) - \eta(\mathbf{X}')\right) < 0\right] \\
&\leq 2C\|\eta^{\text{proj}} - \eta\|_\infty \mathcal{E}(\eta^{\text{proj}})^{\frac{\alpha}{1+\alpha}} \quad \text{(by Condition 3)}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathcal{E}(\eta^{\text{proj}}) &\leq \left(2C\|\eta^{\text{proj}} - \eta\|_\infty\right)^{1+\alpha} \\
&\leq (2C)^{1+\alpha}\left(\sup_{\mathbf{x}} \sum_{j\in S^\star}\sum_{k\geq M+1} \theta_{jk}\phi_k(x_j)\right)^{1+\alpha} \\
&\leq (2BC)^{1+\alpha}\left(\sum_{j\in S^\star}\sum_{k\geq M+1} |\theta_{jk}|\right)^{1+\alpha} \\
&\leq \left(2BC\sqrt{\kappa|S^\star|_0}\right)^{1+\alpha}(1+M)^{-\tau(1+\alpha)},
\end{aligned}$$

by construction of the ϕ_k's.
Combining (15) with what precedes, we get

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le 2\left\{\frac{3}{2}\left(2BC\sqrt{\kappa|S^\star|_0}\right)^{1+\alpha}(1+M)^{-\tau(1+\alpha)} + \left(\frac{4\mathcal{C}\delta}{n}\right)^{1+\alpha}\right.\right.$$
$$\left.\left. + \frac{|S^\star|_0 M \log(1/\beta)}{\delta} + \frac{2\log(2/\varepsilon) + 2K' + |S^\star|_0 \log(2c\delta)}{\delta}\right\}\right] \ge 1-\varepsilon,$$

The function $t \mapsto 3/2\left(2BC\sqrt{\kappa|S^\star|_0}\right)^{1+\alpha}(1+t)^{-\tau(1+\alpha)} + |S^\star|_0 \log(1/\beta)t/\delta$ is upper-bounded at the point

$$t = \left(\frac{2|S^\star|_0\log(1/\beta)}{3\tau(1+\alpha)\left(2BC\sqrt{\kappa|S^\star|_0}\right)^{1+\alpha}\delta}\right)^{-\frac{1}{1+\tau(1+\alpha)}} - 1,$$

so we may write

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le 2\left\{\zeta\delta^{\frac{-\tau(1+\alpha)}{1+\tau(1+\alpha)}} + \left(\frac{4\mathcal{C}\delta}{n}\right)^{1+\alpha} + \frac{2\log(2/\varepsilon) + 2K' + |S^\star|_0 \log(2c\delta)}{\delta}\right\}\right] \ge 1-\varepsilon,$$

where

$$\zeta = \left(3/2\left(2BC\sqrt{\kappa|S^\star|_0}\right)^{1+\alpha}\right)^{\frac{1}{1+\tau(1+\alpha)}}\left(\log(1/\beta)|S^\star|_0\right)^{\frac{\tau(1+\alpha)}{1+\tau(1+\alpha)}}$$
$$\times\left((\tau(1+\alpha))^{-\frac{\tau(1+\alpha)}{1+\tau(1+\alpha)}} + (\tau(1+\alpha))^{\frac{1}{1+\tau(1+\alpha)}}\right).$$

The function $t \mapsto \zeta t^{-\frac{\tau}{1+\tau(1+\alpha)}} + \left(\frac{4\mathcal{C}t}{n}\right)^{1+\alpha}$ (it is sufficient to consider this part) is upper-bounded at the point

$$t = \left(\frac{\zeta\tau(1+\alpha)}{1+\tau(1+\alpha)}\right)^{\frac{1+\tau(1+\alpha)}{(1+\tau(2+\alpha))(1+\alpha)}}\left(\frac{n}{4\mathcal{C}}\right)^{\frac{1+\tau(1+\alpha)}{1+\tau(2+\alpha)}},$$

hence, using the notation

$$\Upsilon = \left(\frac{\zeta\tau(1+\alpha)}{1+\tau(1+\alpha)}\right)^{\frac{1+\tau(1+\alpha)}{(1+\tau(2+\alpha))(1+\alpha)}}\left(\frac{1}{4\mathcal{C}}\right)^{\frac{1+\tau(1+\alpha)}{1+\tau(2+\alpha)}},$$

we obtain that

$$\mathbb{P}\left[\mathcal{E}(\hat{s}) \le 2\left\{\left(\zeta\Upsilon^{-\frac{\tau(1+\alpha)}{1+\tau(1+\alpha)}} + (4\mathcal{C}\Upsilon)^{1+\alpha}\right)n^{\frac{-\tau(1+\alpha)}{1+(2+\alpha)\tau}}\right.\right.$$
$$\left.\left. + \Upsilon^{-1}n^{\frac{-(1+\tau(1+\alpha))}{1+(2+\alpha)\tau}}\left(2\log(2/\varepsilon) + 2K' + |S^\star|_0 \log(2c\Upsilon n^{\frac{1+\tau(1+\alpha)}{1+\tau(2+\alpha)}})\right)\right\}\right] \ge 1-\varepsilon,$$

which concludes the proof.

□

References

AGARWAL, S., GRAEPEL, T., HERBRICH, R., HAR-PELED, S. and ROTH, D. (2005). Generalization bounds for the Area Under the ROC Curve. *Journal of Machine Learning Research*, **6** 393–425. 2

ALQUIER, P. (2006). *Transductive and Inductive Adaptive Inference for Regression and Density Estimation*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 3

ALQUIER, P. (2008). PAC-Bayesian Bounds for Randomized Empirical Risk Minimizers. *Mathematical Methods of Statistics*, **17** 279–304. 3

ALQUIER, P. and BIAU, G. (2013). Sparse Single-Index Model. *Journal of Machine Learning Research*, **14** 243–280. 3

ALQUIER, P. and LOUNICI, K. (2011). PAC-Bayesian Theorems for Sparse Regression Estimation with Exponential Weights. *Electronic Journal of Statistics*, **5** 127–145. 3

AUDIBERT, J.-Y. (2004a). Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincaré : Probabilités et Statistiques*, **40** 685–736. 3

AUDIBERT, J.-Y. (2004b). *Théorie statistique de l'apprentissage : une approche PAC-Bayésienne*. Ph.D. thesis, Université Pierre & Marie Curie - Paris VI. 3

AUDIBERT, J.-Y. and CATONI, O. (2010). Robust linear regression through PAC-Bayesian truncation. Preprint, URL `http://arxiv.org/abs/1010.0072`. 3

AUDIBERT, J.-Y. and CATONI, O. (2011). Robust linear least squares regression. *The Annals of Statistics*, **39** 2766–2794. 3

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *The Annals of Statistics*, **35** 608–633. 8

BASKIOTIS, N., CLÉMENÇON, S., DEPECKER, M. and VAYATIS, N. (2010). TreeRank: an R package for bipartite ranking. In *Proceedings of SMDTA 2010 - Stochastic Modeling Techniques and Data Analysis International Conference*. 18

BOUCHERON, S., BOUSQUET, O. and LUGOSI, G. (2005). Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, **9** 323–375. 9

CARLIN, B. P. and CHIB, S. (1995). Bayesian Model choice via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, Series B*, **57** 473–484. 3

CATONI, O. (2004). *Statistical Learning Theory and Stochastic Optimization*. École d'Été de Probabilités de Saint-Flour XXXI – 2001, Springer. 3, 7, 20

CATONI, O. (2007). *PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning*, vol. 56 of *Lecture notes – Monograph Series*. Institute of Mathematical Statistics. 3

CLÉMENÇON, S., LUGOSI, G. and VAYATIS, N. (2008). Ranking and empirical risk minimization of U-statistics. *The Annals of Statistics*, **36** 844–874. 2, 5, 9, 27

CLÉMENÇON, S. and ROBBIANO, S. (2011). Minimax learning rates for bipartite ranking and plug-in rules. In *Proceedings of the 28th International Conference on Machine Learning*. 441–448. 2, 10, 12, 18, 29

CLÉMENÇON, S. and VAYATIS, N. (2009). Tree-based ranking methods. *IEEE Transactions on Information Theory*, **55** 4316–4336. 2

CLÉMENÇON, S., DEPECKER, M. and VAYATIS, N. (2011). Adaptive partitioning schemes for bipartite ranking adaptive partitioning schemes for bipartite ranking: How to grow and prune a ranking tree. *Machine Learning*, **83** 31–69. 18

CSISZÁR, I. (1975). i-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, **3** 146–158. 20

DALALYAN, A. S. and SALMON, J. (2012). Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, **40** 2327–2355. 3

DALALYAN, A. S. and TSYBAKOV, A. B. (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity. *Machine Learning*, **72** 39–61. 3, 6

DALALYAN, A. S. and TSYBAKOV, A. B. (2012). Sparse Regression Learning by Aggregation and Langevin Monte-Carlo. *Journal of Computer and System Sciences*, **78** 1423–1443. 3

FREUND, Y., IYER, R. D., SCHAPIRE, R. E. and SINGER, Y. (2003). An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research*, **4** 933–969. 2, 18

GREEN, D. M. and SWETS, J. A. (1966). *Signal detection theory and psychophysics*. Wiley. 2

GUEDJ, B. (2013). *pacbpred: PAC-Bayesian Estimation and Prediction in Sparse Additive Models*. R package version 0.92.2, URL http://cran.r-project.org/web/packages/pacbpred/index.html. 3, 13

GUEDJ, B. and ALQUIER, P. (2013). PAC-Bayesian estimation and prediction in sparse additive models. *Electronic Journal of Statistics*, **7** 264–291. 3, 5, 13

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun Stochastic Search for "Large p" Regression. *Journal of the American Statistical Association*, **102** 507–516. 3

HASTIE, T. and TIBSHIRANI, R. (1986). Generalized Additive Models. *Statistical Science*, **1** 297–318. 6

LECUÉ, G. (2006). Optimal oracle inequality for aggregation of classifiers under low noise condition. In *Proceedings of the 19th Conference on Learning Theory*. 364–378. 9

LEUNG, G. and BARRON, A. R. (2006). Information theory and mixing least-squares regressions. *IEEE Transactions on Information Theory*, **52** 3396–3410. 8

LI, C., JIANG, W. and TANNER, M. (2013). General oracle inequalities for Gibbs posterior with application to ranking. In *Proceedings of the 26th Conference on Learning Theory*. 512–521. 2, 5

MAMMEN, E. and TSYBAKOV, A. B. (1999). Smooth discrimination analysis. *The Annals of Statistics*, **27** 1808–1829. 9

MCALLESTER, D. A. (1999). Some PAC-Bayesian Theorems. *Machine Learning*, **37** 355–363. 2

PETRALIAS, A. and DELLAPORTAS, P. (2012). An MCMC model search algorithm for regression problems. *Journal of Statistical Computation and Simulation*, **0** 1–19. 3, 13

RAKOTOMAMONJY, A. (2004). Optimizing area under ROC curve with SVMs. In *Proceedings of European Conference on Artificial Intelligence Workshop on ROC Curve and AI*. 2

RIDGWAY, J., ALQUIER, P., CHOPIN, N. and LIANG, F. (2014). PAC-Bayesian AUC classification and scoring. In *Advances in Neural Information Processing Systems 27 (NIPS)*. 658–666. 3, 8

RIGOLLET, P. and TSYBAKOV, A. B. (2012). Sparse Estimation by Exponential Weighting. *Statistical Science*, **27** 558—575. 6

ROBBIANO, S. (2013). Upper bounds and aggregation in bipartite ranking. *Electronic Journal of Statistics*, **7** 1249–1271. 2, 9, 11

SERFLING, R. J. (1980). *Approximation theorems of mathematical statistics*. Wiley. 22

SHAWE-TAYLOR, J. and WILLIAMSON, R. C. (1997). A PAC analysis of a Bayes estimator. In *Proceedings of the 10th annual conference on Computational Learning Theory*. ACM, 2–9. 2

STONE, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13** 689–705. 6

SUZUKI, T. (2012). PAC-Bayesian Bound for Gaussian Process Regression and Multiple Kernel Additive Model. In *Proceedings of the 25th annual conference on Computational Learning Theory*. 1–20. 3

TSYBAKOV, A. B. (2004). Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, **32** 135–166. 9

TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Statistics, Springer. 11