A data-dependent generalisation error bound for the AUC

Nicolas Usunier usunier@poleia.lip6.fr
Massih-Reza Amini Amini@poleia.lip6.fr
Patrick Gallinari gallinari@poleia.lip6.fr
Laboratoire d'Informatique de Paris 6
8, rue du Capitaine Scott
75015 Paris, France

Abstract

The optimisation of the Area Under the ROC Curve (AUC) has recently been proposed for learning ranking functions. However, the estimation of the AUC of a function on the true distribution of the examples based on its empirical value is still an open problem. In this paper, we present a *data-dependent* generalisation error bound for the AUC. This bound presents the advantage to be tight, but it also allows to draw practical conclusions on learning algorithms which optimise the AUC. In particular, we show that in the case of AUC, kernel function classes have strong generalisation guarantees provided that the weights of the functions are small, suggesting that regularisation procedures which tend to limit the norm of the weight vector may lead to better generalisation performance for algorithms which optimise the AUC.

1. Introduction

Many supervised machine learning (ML) applications are a bipartite ranking problem, where the goal is to learn a scoring function which gives higher scores to *positive* examples than to *negative* ones. For example, in metasearch, machine learning can be used to combine the outputs of search engines in order to improve the ranks of the relevant elements (Aslam & Montague, 2001). Another example is the task of automatic text summarization by extraction, where a summarization system takes as input a document, and provides an ordered list of some of the document sentences, in which the top-ranked ones should reflect the main idea of its content (Kupiec et al., 1995; Amini & Gallinari, 2002; Amini et al., 2005).

While these ranking tasks were originally dealt with by learning a classifier, it has recently been shown that the error rate of a classifier is lowly correlated to its Area Under the ROC Curve (AUC) (Caruana & Niculescu-Mizil, 2004; Cortes & Mohri, 2004), a measure of the ranking performance of a function, equal to the probability on a given set that a positive instance has a greater score than a negative one. This observation has led to a new kind of learning algorithms, designed specifically to optimise the AUC (Yan et al., 2003; Rakotomamonjy, 2004; Herschtal & Raskutti, 2004). If we are given a training set S composed of n positive instances $(x_i)_{i=1}^n$ and m negative instances $(x'_j)_{j=1}^m$, the AUC of a function f is equal to (see e.g. (Cortes & Mohri, 2004)):

$$AUC(f,S) = \frac{1}{nm}\sum_{i,j}(I_{f(x_i)>f(x'_j)} + \frac{1}{2}I_{f(x_i)=f(x'_j)})$$
$$\geq 1 - \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{f(x_i)\leq f(x'_j)}$$

where $I_{pr} = 1$ if pr is true and 0 otherwise. The optimisation of the AUC carried out by these specific algorithms is to find a function f in a given class $\mathcal{F}$ which minimises:

$$\hat{\mathbb{E}}_S L_f = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} L(f(x_i) - f(x'_j)) \qquad (1)$$

where $L : (x, x') \mapsto L(f(x) - f(x'))$ is a loss function, approximating or upper bounding $I_{f(x)\leq f(x')}$ (see e.g. (Freund et al., 2003; Herschtal & Raskutti, 2004).

In this paper, we are interested in the generalisation properties of the AUC. Formally, assuming that the sample S is drawn i.i.d. according to an unknown distribution $\mathcal{D}$, we want to bound the expected AUC

Appearing in *Proceedings of the ICML 2005 workshop on ROC Analysis in Machine Learning*, Bonn, Germany, 2005.

of f based on its AUC on the training set and the class of functions $\mathcal{F}$, by the standard uniform bound:

$$\mathbb{E}_{x,x'} L_f - \hat{\mathbb{E}}_S L_f \leq \sup_{f \in \mathcal{F}} (\mathbb{E}_{x,x'} L_f - \hat{\mathbb{E}}_S L_f) \qquad (2)$$

where the expectation $\mathbb{E}_{x,x'}$ is taken over the conditional distributions $\mathcal{D}$ given the class labels (x is a positive and x' a negative instance).

Such bounds are a key characteristic of the optimisation of the AUC from a ML point of vue. Indeed, if the right hand side of the above inequality tends to 0 when the sample size tends to infinity, the learning process is then consistent (Vapnik, 2000). Finite sample bounds which give an indication on the convergence rate of the learning procedure, are also of great interest. Indeed, such bounds allow to show that optimising different classes of functions may have different generalisation guarantees, and therefore will influence the design of the learning algorithms as well as the choice of the algorithms in practical applications.

We present here a new approach for the computation of *data-dependent* bounds for the AUC, inspired by the work of (Bartlett & Mendelson, 2003) on data-dependent bounds for classification. The main benefit of such approaches is that they do not make any assumptions on the nature of the distribution $\mathcal{D}$, while the bounds can be calculated based on the training data and therefore may provide tighter bounds than the ones based on distribution-free estimates of the complexity of the class of functions like the VC-dimension.

With the intention of obtaining tight generalisation error bounds for the AUC, we obtain a new data-dependent generalisation error bound, and show a particular instantiation on kernel function classes with bounded weight vectors. Our bounds confirm the previous generalisation error bounds for the AUC of (Agarwal et al., 2005; Freund et al., 2003) in that the convergence rate is mainly controlled by the number of instances of the minority class for unbalanced datasets. However, our result also leads to the conclusion that, for this class of functions, the generalisation error will be small if the weights are small, independently from the dimension of the (implicit) feature space. This suggests that the algorithms which optimise the AUC while controlling the size of the weights, like the SVM for the AUC (Rakotomamonjy, 2004), will have good generalisation guarantees. It is an extension to the case of AUC of the same observation in classification, which is that controlling the size of the weights in a learning algorithm may lead to better generalisation performance, like in the well-known SVM (Vapnik, 1998; Shawe-Taylor & Critiani, 2004).

The remainder of the paper is organised as follows. In section 2, we present some related work on the generalisation properties of the AUC. Then in section 3, we define the quantity which will be used as function class complexity for the AUC and prove that it allows to bound the generalisation error. And finally we give a data-dependent bound for the AUC in the special case of kernel classes in section 4.

2. Related work

Generalisation error bounds for the AUC have already been proved using distribution-free estimates of the class complexity (Freund et al., 2003; Agarwal et al., 2005). These bounds depend on the number of positive and negative instances in the training set and show in particularly that for highly unbalanced datasets, the rate of convergence will mainly be controlled by the number of examples in the minority class. (Freund et al., 2003) proposed a bound using the VC dimension of the class of functions, and (Agarwal et al., 2005) defined a new function class complexity, called the bipartite rank-shatter coefficients. With this complexity, they found tighter bounds for the linear ranking function classes than the ones in (Freund et al., 2003). However, these bipartite rank-shatter coefficients are difficult to evaluate for other classes of functions than linear or polynomial ones. Moreover, in these special cases, the rank-shatter coefficients of $\mathcal{F}$ depend on the dimension d of the feature space, making it too loose for function classes in a large (implicit) feature space like kernel machines. In our approach, we will prove bounds which, when the class of functions $\mathcal{F}$ is linear with a bounded weight vector, have a convergence rate of $\bigcirc(\sqrt{\frac{n+m}{nm}})$, which is the same as the one found by (Agarwal et al., 2005). However the main difference with their approach is that the bound we propose does not depend on the dimension of the feature space, making it particularly convenient with kernel functions. Our approach can be seen as the analogous of the Rademacher complexity for classification, a powerful tool used for providing tight data-dependent bounds for the kernel machines used in classification (Bartlett & Mendelson, 2003).

Recently, (Clémençon et al., 2005) presented an in-depth statistical analysis of the generalisation properties of a quantity they call the *ranking risk*, which differs from the AUC mainly by a constant factor. Therefore, their asymptotic results on the ranking risk can be applied to the case of AUC. However, the constant factor between the AUC and the ranking risk depends on the true class probabilities, which are unknown. Therefore, the use of their results to obtain tight finite

sample bounds is not straightforward.

In order to obtain tight, data-dependent convergence rates, we propose another approach, specific to the problem of AUC. We do not claim that our approach is optimal, but it has the advantage of easily providing bounds on the convergence rate without considering the real class probability distributions.

3. A new generalisation error bound

3.1. Notations and Definitions

From now on, we suppose that the class labels are $\{1, -1\}$, that there is a mapping from the examples to a feature space $\mathcal{X}$, and that the example follow an unknown distribution $\mathcal{D}$ over $\mathcal{X} \times \{1, -1\}$. We moreover denote by $\mathcal{D}_1$ and $\mathcal{D}_{-1}$ the conditional distributions of $\mathcal{D}$ given the class label (1 and -1 respectively), and $\mathcal{D}_1^n$ and $\mathcal{D}_{-1}^m$ the product distribution of $\mathcal{D}_1$ and $\mathcal{D}_{-1}$ over $\mathcal{X}^n$ and $\mathcal{X}^m$ respectively. Finally, we will denote by $S = (x_1, ..., x_n, x'_1, ..., x'_m)$ a sample set drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$, which means that all x_i and $x'j$ are independent and that the x_is are positive instances drawn iid according to $\mathcal{D}_1$ and the x'_js are negative instances drawn i.i.d. according to $\mathcal{D}_{-1}$[1].

From equation 1, the optimisation of the AUC can be carried out by minimising, for some loss function L and f in a given class $\mathcal{F}$, the quantity $\frac{1}{nm}\sum_{i=1}^m \sum_{j=1}^m L(f(x_i) - f(x'_j))$.

More generally, we will consider a class of functions $\mathcal{Q}$ mapping from $\mathcal{X}^2$ to $[0, 1]$. A typical example is $\mathcal{Q} = \{Q_f : (x, x') \mapsto I_{f(x) \leq f(x')}, f \in \mathcal{F}\}$. Let us introduce :

$$\hat{\mathbb{E}}_S Q = \frac{1}{nm} \sum_{(i,j)} Q(x_i, x'_j)$$

and

$$\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q = \mathbb{E}_{x \sim \mathcal{D}_1, x' \sim \mathcal{D}_{-1}} Q(x, x')$$

where $x \sim \mathcal{D}_1$ means that the random variable x follows distribution $\mathcal{D}_1$. Our goal is to give an upper bound on (see equation 2):

$$\sup_{Q \in \mathcal{Q}} (\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q - \hat{\mathbb{E}} Q) \tag{3}$$

In order to obtain a data-dependent bound, we will make use of Rademacher variables and Rademacher averages (see e.g. (Bartlett & Mendelson, 2003)). A random variable σ is a Rademacher variable if $P(\sigma = 1) = P(\sigma = -1) = \frac{1}{2}$, and we will make use of Rademacher averages of the following form:

$$\mathbb{E}_\sigma \sup_{\theta \in \Theta} \sum_{k=1}^N \sigma_k h_k(\theta)$$

where $\sigma_1, ..., \sigma_N$ are independent Rademacher variables, and, for all k, h_k is a real-valued function. Such Rademacher averages have been used as the complexity of the class of function to obtain data-dependent bounds for classification (see e.g. (Bartlett & Mendelson, 2003)), where Θ was the class of functions and $h_k(\theta) = \theta(x_k)$ for an example x_k.

Let us define the following Rademacher averages, which we will use as a function class complexity in our results:

$$\hat{R}_{n,m}^{AUC}(\mathcal{Q}) = 4\mathbb{E}_{\sigma,\nu} \sup_{Q \in \mathcal{Q}} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \frac{\sigma_i + \nu_j}{2} Q(x_i, x'_j) \tag{4}$$

where $\sigma = (\sigma)_{i=1}^n$, and $\nu = (\nu)_{i=1}^m$ are two independent sequences of independent Rademacher variables. And

$$R_{n,m}^{AUC}(\mathcal{Q}) = \mathbb{E}_S \hat{R}_{n,m}^{AUC}(\mathcal{Q}) \tag{5}$$

where the expectation is taken over all samples S drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$.

The main interest of these averages is that they provide a generalisation error bound for the AUC by bounding (3) using $R_{n,m}^{AUC}(\mathcal{Q})$ (theorem 1). Then, we prove in lemma 4 that we can approximate $R_{n,m}^{AUC}(\mathcal{Q})$ by its empirical value on the training data $\hat{R}_{n,m}^{AUC}(\mathcal{Q})$, which shows that we can obtain data-dependent bounds for the AUC. In theorem 6, we present the calculation of the bound for the specials cases of classes of kernel functions with bounded weights, which will be used to draw our practical conclusions.

3.2. Main result

Theorem 1. *Let $\mathcal{Q}$ be a class of functions mapping $\mathcal{X}^2$ to $[0, 1]$, let $S = (x_1, ..., x_n, x'_1, ..., x'_m)$ be a sample of size $n + m$ drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$. Then, with probability $1 - \delta$, all Q in $\mathcal{Q}$ satisfy:*

$$\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q \leq \hat{\mathbb{E}} Q + R_{n,m}^{AUC}(\mathcal{Q}) + \sqrt{\frac{(n+m)}{2nm} \ln(1/\delta)}$$

Proof. for all $Q \in \mathcal{Q}$, we have:

$$\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q - \hat{\mathbb{E}}_S Q \leq \sup_{Q \in \mathcal{Q}} (\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q - \hat{\mathbb{E}}_S Q) \tag{6}$$

[1] All our results will be stated for samples drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$ in order to simplify the notations. However, since the order of the instances of the training set is of no importance, all our proofs could be conducted like in (Agarwal et al., 2005), considering samples of size $n + m$ drawn according to $\mathcal{D}^{n+m}$, but conditionally on the label of each instance. Therefore, our approach is not less general than theirs.

The main idea of the whole proof is to use a symmetrisation procedure. We will obtain an upper bound on the right term of 6 which is symmetric in two samples of the same size, the sample S and another arbitrary sample $\tilde{S}$ taken according to the same distribution (equation 7). This symmetric expression will enable us to introduce Rademacher variables which correspond to random permutations of examples from S to $\tilde{S}$ (lemma 3) which will allow us to introduce the Rademacher average we defined in equation 5.

The first step of the proof is to bound the right hand side of the inequality (6) using the results of the following lemma, which is due to a theorem by McDiarmid (McDiarmid, 1989) (For clarity in the presentation, the proof is deferred to Appendix A):

Lemma 2. *Let $\mathcal{Q}$ be a class of functions mapping from $\mathcal{X}^2$ to $[0,1]$. Then, with probability $1-\delta$ over all samples S drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$, we have:*

$$\sup_{Q\in\mathcal{Q}}(\mathbb{E}_{\mathcal{D}_1\times\mathcal{D}_{-1}}Q - \hat{\mathbb{E}}_S Q) \leq \sqrt{\frac{(n+m)}{2nm}\ln(1/\delta)} + \mathbb{E}_{S\sim\mathcal{D}_1^n\times\mathcal{D}_{-1}^m}\sup_{Q\in\mathcal{Q}}(\mathbb{E}_{\mathcal{D}_1\times\mathcal{D}_{-1}}Q - \hat{\mathbb{E}}_S Q)$$

Let us now consider a second sample $\tilde{S} = (\tilde{x}_1, ..., \tilde{x_n}, \tilde{x}'_1, ..., \tilde{s'_m})$ of the same size and drawn according to the same distribution as S. We will now denote $\mathbb{E}_{S\sim\mathcal{D}_1^n\times\mathcal{D}_{-1}^m}$ as $\mathbb{E}_S$ and $\mathbb{E}_{\tilde{S}\sim\mathcal{D}_1^n\times\mathcal{D}_{-1}^m}$ as $\mathbb{E}_{\tilde{S}}$ when the context is clear. It is easy to see that we have $\mathbb{E}_{\tilde{S}}\hat{\mathbb{E}}_{\tilde{S}}Q = \mathbb{E}_{\mathcal{D}_1\times\mathcal{D}_{-1}}Q$, and, as a consequence:

$$\begin{aligned}\mathbb{E}_S \sup_{Q\in\mathcal{Q}} \quad & (\mathbb{E}_{\mathcal{D}_1\times\mathcal{D}_{-1}}Q - \hat{\mathbb{E}}_S Q) \\ & = \mathbb{E}_S \sup_{Q\in\mathcal{Q}}(\mathbb{E}_{\tilde{S}}\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q) \\ & = \mathbb{E}_S \sup_{Q\in\mathcal{Q}}\mathbb{E}_{\tilde{S}}[\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q] \\ & \leq \mathbb{E}_{S,\tilde{S}}\sup_{Q\in\mathcal{Q}}(\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q)\end{aligned} \tag{7}$$

where the last inequality is obtained after remarking that we have, for all Q and S:

$$\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q \leq \sup_{Q'\in\mathcal{Q}} \hat{\mathbb{E}}_{\tilde{S}}Q' - \hat{\mathbb{E}}_S Q'$$

and then taking the expectation over $\tilde{S}$ and the supremum over $\mathcal{Q}$.

Using equation 7, which is symmetric in S and $\tilde{S}$, we introduce the Rademacher variables through the following lemma (the full proof is given in Appendix B):

Lemma 3. *With S and $\tilde{S}$ defined as above, and considering $\sigma = (\sigma_i)_{i=1}^n$ and $\nu = (\nu_j)_{j=1}^m$, two sequences of independent Rademacher variables of size n and m respectively, the two following quantities are equal:*

1. $\mathbb{E}_{S,\tilde{S}}\sup_{Q\in\mathcal{Q}}(\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q)$
2. $\mathbb{E}_{\sigma,\nu,S,\tilde{S}}\sup_{Q\in\mathcal{Q}}\{\frac{1}{nm}\sum_{i=1}^n\sum_{j=1}^m \frac{\sigma_i+\nu_j}{2}Q(\tilde{x}_i, \tilde{x}'_j) + \frac{\sigma_i-\nu_j}{2}Q(\tilde{x}_i, x'_j) - \frac{\sigma_i-\nu_j}{2}Q(x_i, \tilde{x}'_j) - \frac{\sigma_i+\nu_j}{2}Q(x_i, x'_j))\}$

Proof. (of lemma 3) consider the part of the second expression of the lemma which is inside the supremum, and, for a given i, set $\sigma_i = -1$. Then all the sum is the same as in the first expression if we had swapped the examples x_i and $\tilde{x}_i$ from the two considered samples S and $\tilde{S}$. More generally, for any instance of σ and ν, this part of the expression corresponds to a swap between S and $\tilde{S}$ of the i-th x_is and the j-th x'_js for which $\sigma_i = -1$ and $\nu_j = -1$. Since S and $\tilde{S}$ have the same distribution, swapping elements from one sample to the other does not change the expectation over $S, \tilde{S}$. Therefore, for all σ and ν, the expectation over $S, \tilde{S}$ of the supremum of the second expression of the lemma is equal to the first expression, and the theorem follows by taking the expectation over σ, ν. □

Back to the demonstration of theorem 1 and decomposing the second expression of lemma 3 into four independent expectations of supremums of Rademacher processes, we can notice that σ has the same distribution as $-\sigma$, and that it is the same for ν and $-\nu$, $(x_i)_{i=1}^n$ and $(\tilde{x}_i)_{i=1}^n$ and finally $(x'_j)_{j=1}^m$ and $(\tilde{x}'_j)_{j=1}^m$, and that therefore the four terms obtained have the same value. Thus, we can claim that:

$$\mathbb{E}_{S,\tilde{S}}\sup_{Q\in\mathcal{Q}}(\hat{\mathbb{E}}_{\tilde{S}}Q - \hat{\mathbb{E}}_S Q) \leq R_{n,m}^{AUC}(\mathcal{Q})$$

where $R_{n,m}^{AUC}(\mathcal{Q})$ is defined in equation 5. Putting it together with equation 7 and with lemma 2 gives theorem 1. □

4. Calculating the bound using the data

An important characteristic of $R_{n,m}^{AUC}(\mathcal{Q})$ is that it can be approximated using the training data. Indeed, we have the following lemma, which, together with theorem 1 give a generalisation error bound for the AUC that can be computed on the data.

Lemma 4. *Let $\mathcal{Q}$ be a class of functions mapping $\mathcal{X}^2$ to $[0,1]$. Let $S = (x_1, ..., x_n, x'_1, ..., x'_m)$ be a sample drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$, then, with probability at least $1-\delta$, we have:*

$$R_{n,m}^{AUC}(\mathcal{Q}) \leq \hat{R}_{n,m}^{AUC}(\mathcal{Q}) + 4\sqrt{\frac{n+m}{2nm}\ln(1/\delta)}$$

Proof. The proof is analogous to the proof of lemma 2 (see Appendix A), by applying McDiarmid's theorem

to the following function:

$$f(S) = \mathrm{E}_{\sigma,\nu} \sup_{Q \in \mathcal{Q}} \frac{1}{nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{\sigma_i + \nu_j}{2} Q(x_i, x'_j)$$

and by noting that for all x_k and $\hat{x}_k$, all $Q \in \mathcal{Q}$ and all σ, ν we have:

$$\frac{1}{nm} \mid \sum_{j=1}^{m} \frac{\sigma_k + \nu_j}{2} Q(x_k, x'_j) - \sum_{j=1}^{m} \frac{\sigma_k + \nu_j}{2} Q(\hat{x}_k, x'_j) \mid \leq \frac{1}{n}$$

and, for all x'_k and $\hat{x}'_k$ we have, for all $Q \in \mathcal{Q}$, all σ, ν:

$$\frac{1}{nm} \mid \sum_{i=1}^{n} \frac{\sigma_i + \nu_k}{2} Q(x_i, x'_k) - \sum_{i=1}^{n} \frac{\sigma_i + \nu_k}{2} Q(x_i, \hat{x}'_k) \mid \leq \frac{1}{m}$$

□

Although we now know that there exist some data-dependent bounds for the AUC, we need to explicitly calculate $\hat{R}^{AUC}_{n,m}(\mathcal{Q})$ for some specific $\mathcal{Q}$ in order to draw practical conclusions. In order to do this, consider a kernel K over $\mathcal{X}^2$, B a strictly positive real number and the following class of functions:

$$\mathcal{F}_{K,B} = \{f_w : \mathcal{X} \to \mathrm{R}, f_w(x) = K(w, x) \mid ||w||_K \leq B\} \tag{8}$$

where $||x||_K = \sqrt{K(x,x)}$. Given a sample $S = (x_1, ..., x_n, x'_1, ..., x'_m)$, the goal here is to provide a generalisation error bound in terms of the variables $\zeta_{ij} = \left[\rho - K(w, x_i) - K(w, x'_j)\right]_+$, where $[x]_+$ denotes the positive part of x and ρ plays an analogous role for the AUC as the margin for classification. It is to be noted that the ζ_{ij} are the slack variables for the SVM which optimise the AUC (Rakotomamonjy, 2004). In order to simplify the notations we suppose here that $\rho = 1$, but similar results stand for any reasonable values of ρ.

In order to make the calculations, we need to define $\phi : \mathbb{R} \to [0, 1]$, the 1-Lipschitz function such that: $\phi(x) = 0$ if $x \geq 1$, $\phi(x) = 1$ if $x \leq 0$ and $\phi(x) = 1 - x$ for $0 \leq x \leq 1$. Then, we have, for f in $\mathcal{F}_{K,B}$:

$$I_{f(x_i) \leq f(x'_j)} \leq \phi(f(x_i) - f(x'_j)) \leq \zeta_{ij} \tag{9}$$

and therefore $\mathrm{E}_{x,x'} I_{f(x) \leq f(x')} \leq \mathrm{E}_{x,x'} \phi(f(x) - f(x'))$. Applying theorem 1 and lemma 4 to the right hand term of this inequality leads to the fact that, with probability at least $1 - \delta$ over the samples S drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$:

$$\mathrm{E}_{x,x'} I_{f(x) \leq f(x')} \leq \mathrm{E}_{x,x'} \phi(f(x) - f(x')) + \hat{R}^{AUC}_{n,m}(\phi \circ \mathcal{F}_{K,B}) + 5\sqrt{\frac{(n+m)}{2nm} \ln(2/\delta)} \tag{10}$$

where we have used the abuse of notation $\phi \circ \mathcal{F}_{K,B} = \{(x, x') \mapsto \phi(f(x) - f(x')) \mid f \in \mathcal{F}_{K,B}\}$. This expression leads to the following lemma which provides an upper bound for the generalisation error that can be computed on the training set.

Lemma 5. *Let ψ be Lipshitz with constant L. Then, for all $S = (x_1, ..., x_n, x'_1, ..., x'_m) \in \mathcal{X}^{n+m}$, we have:*

$$\hat{R}^{AUC}_{n,m}(\psi \circ \mathcal{F}_{K,B}) \leq A_{n,m,B,L} \sqrt{\sum_{(i,j)} ||x_i||^2_K + ||x'_j||^2_K - 2K(x_i, x'_j)}$$

with $A_{n,m,B,L} = \frac{2LB\sqrt{2(n+m)}}{nm}$.

Proof. From the definition of $\hat{R}^{AUC}_{n,m}$, we have:

$$\begin{aligned}
&\hat{R}^{AUC}_{n,m}(\psi \circ \mathcal{F}_{K,B}) \\
&\quad = 4\mathrm{E}_{\sigma,\nu} \sup_{||w||_K \leq B} \frac{1}{nm} \sum_{(i,j)} \frac{\sigma_i + \nu_j}{2} \psi(K(w, x_i) - K(w, x'_j)) \\
&\quad \leq \frac{2}{nm} \sum_{j=1}^{m} \mathrm{E}_\sigma \sup_{||w||_K \leq B} \sum_{i=1}^{n} \sigma_i \psi(K(w, x_i) - K(w, x'_j)) \\
&\qquad + \frac{2}{nm} \sum_{i=1}^{n} \mathrm{E}_\nu \sup_{||w||_K \leq B} \sum_{j=1}^{m} \nu_j \psi(K(w, x_i) - K(w, x'_j))
\end{aligned} \tag{11}$$

We obtain therefore a sum of m Rademacher averages (RA) of size n and n RA of size m. Now denoting $<,>$ the scalar product in the implicit space of K and φ the projection such that $K(x, x') = < \varphi(x), \varphi(x') >$, and, for an element of the implicit space y, $||y|| = \sqrt{< y, y >}$, we can apply theorem 7 of (Meir & Zhang, 2003) (because ψ is L-Lipschitz) to the previous Rademacher averages,from which we directly get:

$$\begin{aligned}
&\hat{R}^{AUC}_{n,m}(\psi \circ \mathcal{F}_{K,B}) \\
&\quad \leq \frac{2L}{nm} \sum_{j=1}^{m} \mathrm{E}_\sigma \sup_{||w||_K \leq B} \sum_{i=1}^{n} \sigma_i [K(w, x_i) - K(w, x'_j)] \\
&\qquad + \frac{2L}{nm} \sum_{i=1}^{n} \mathrm{E}_\nu \sup_{||w||_K \leq B} \sum_{j=1}^{m} \nu_j [K(w, x_i) - K(w, x'_j)] \\
&\quad \leq \frac{2LB}{nm} \sum_{j=1}^{m} \mathrm{E}_\sigma || \sum_{i=1}^{n} \sigma_i [\varphi(x_i) - \varphi(x'_j)] || \\
&\qquad + \frac{2LB}{nm} \sum_{i=1}^{n} \mathrm{E}_\nu || \sum_{j=1}^{m} \nu_j [\varphi(x_i) - \varphi(x'_j)|| \\
&\quad \leq \frac{2LB(n+m)}{nm} \{ \frac{1}{n+m} \sum_{j=1}^{m} \mathrm{E}_\sigma || \sum_{i=1}^{n} \sigma_i [\varphi(x_i) - \varphi(x'_j)] ||^2 \\
&\qquad + \frac{1}{n+m} \sum_{i=1}^{n} \mathrm{E}_\nu || \sum_{j=1}^{m} \nu_j [\varphi(x_i) - \varphi(x'_j)] ||^2 \}^{\frac{1}{2}}
\end{aligned} \tag{12}$$

where the second inequality is obtained using the bilinearity of $<,>$ and Cauchy-Schwartz's inequality, and the third is due to $||y|| = \sqrt{||y||^2}$ and to two consecutive applications of Jensen's inequality (since the square root is concave). Considering each term separately, we have for example:

$$
\begin{aligned}
&||\sum_{i=1}^{n} \sigma_i \sigma_i [\varphi(x_i) - \varphi(x'_j)]||^2 \\
&\quad = < \sum_{i=1}^{n} \sigma_i [\varphi(x_i) - \varphi(x'_j)], \sum_{i=1}^{n} \sigma_i \varphi(x_i) - \varphi(x'_j)] > \\
&\quad = \sum_{i=1}^{n} \sum_{l=1}^{n} \sigma_i \sigma_l < [\varphi(x_i) - \varphi(x'_j)], [\varphi(x_l) - \varphi(x'_j)] >
\end{aligned}
$$

And, since the σ_i and σ_l are independent for $i \neq l$ with 0 mean, only the terms in $\sigma_i \sigma_i ||[\varphi(x_i) - \varphi(x'_j)]||^2$ remain in the last sum when taking the expectation over σ, and, therefore:

$$
\begin{aligned}
&\mathrm{E}_\sigma ||\sum_{i=1}^{n} \sigma_i [\varphi(x_i) - \varphi(x'_j)]||^2 = \\
&\qquad \sum_{i=1}^{n} [||x_i||_K^2 + ||x'_j||_K^2 - 2K(x_i, x'_j)]
\end{aligned} \tag{13}
$$

Using the same reasoning for ν and putting it in equation 12 proves the lemma . □

Finally, using the result of lemma 5 in equation 10 leads to the following theorem:

Theorem 6. *Using the notations defined above, with probability at least $1 - \delta$ over samples $S = (x_1, ..., x_n, x'_1, ..., x'_m)$ drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$, the following inequality holds for all $f \in \mathcal{F}_{K,B}$:*

$$
\begin{aligned}
\mathrm{E}_{x,x'} I_{f(x) \leq f(x')} &\leq \frac{1}{nm} \sum_{(i,j)} \phi(f(x_i) - f(x'_j)) \\
&+ \frac{2B\sqrt{2(n+m)}}{nm} \sqrt{\sum_{(i,j)} [||x_i||_K^2 + ||x'_j||_K^2 - 2K(x_i, x'_j)]} \\
&+ 5\sqrt{\frac{(n+m)}{2nm} \ln(2/\delta)}
\end{aligned}
$$

And, noting $c = \max(\max_i ||x_i||_K, \max_j ||x'_j||_K)$:

$$
\begin{aligned}
\mathrm{E}_{x,x'} I_{f(x) \leq f(x')} &\leq \frac{1}{nm} \sum_{(i,j)} \phi(f(x_i) - f(x'_j)) \\
&+ 4\sqrt{2}Bc\sqrt{\frac{n+m}{nm}} + 5\sqrt{\frac{(n+m)}{2nm} \ln(2/\delta)}
\end{aligned} \tag{14}
$$

We can notice, using equation 9, that the equalities of the theorem are still true if we replace $\sum_{(i,j)} \phi(f(x_i) - f(x'_j))$ with $\sum_{(i,j)} \zeta_{ij}$, where ζ_{ij} are defined above. We therefore proved some generalisation bounds which are analogous to the margin-based, data-dependent bounds for kernel machines (Bartlett & Mendelson, 2003). Using equation 14, it is also easy to see that if the data lie in a ball of the implicit space (i.e. $\forall x \in \mathcal{X}, ||x||_K \leq B'$), then the rate of convergence is $\bigcirc(\sqrt{\frac{n+m}{nm}})$, and depends only on the maximal norm of the weight vector of the functions. In particular, the bound does not depend on the dimension of the feature space, and can be applied to any kernel. As direct consequences, this bound shows at first that optimising the relative difference of scores between instances of different classes, which is the analogous to optimising the margin in classification, actually leads to good performance guarantees. This result is interesting, because RankBoost applied to the AUC (Freund et al., 2003), or the SVM for the AUC (Rakotomamonjy, 2004) optimise these relative differences of scores[2]. Secondly, this bound shows that when we can control the norm of the weight vector, we have some generalisation guarantees which are in particular independent from the dimension of the feature space, implicit or explicit. Therefore, as in classification, sophisticated kernels can be used in order to obtain non-linear ranking functions, and generalisation performance is still guaranteed. More generally, we have shown that the kernels used in classification can also be used to learn a ranking function.

5. Conclusion and perspectives

In this paper, we have shown that data-dependent bounds could be obtained for the AUC, and calculated the bound in the particular case of kernel classes of functions with bounded weights. Our results also confirm that the generalisation error is mainly controlled by the number of instances in the minority class. Moreover, in terms of practical conclusions, we have shown that the relative difference of scores between instances of the two different classes, the analogous of the margin in classification, is closely related to the generalisation performance of the AUC. We also showed that kernel functions can be used to learn complex functions, in particular non-linear functions, while keeping their generalisation performances given that the weights are bounded. This leads to the last interesting conclusion that the size of the weights is closely

[2]It is to be noted here that since the weights in RankBoost are not chosen to be small, this bound is not applicable directly. However, we showed a relationship between the generalisation performance of the optimisation of the AUC and the maximisation of the difference of the relative scores.

related to the generalisation performance, and, in the case studied here, is more important than the dimension of the implicit space.

Finally, the proofs follow the same steps as (Bartlett & Mendelson, 2003) for the data-dependent bounds for classification. However, we had to define specific Rademacher averages for the case of AUC to derive interesting and useful data-dependent bounds. This shows in particular that they can be a useful tool to analyse the generalisation properties of the AUC. However, the actual Rademacher averages used lack the structural results which exist for the Rademacher complexity used in classification (Bartlett & Mendelson, 2003). Therefore, more work is needed to study the complexity defined in the paper, or to define another complexity measure of classes of functions which would allow at the same time for data-dependent bounds and more general results.

6. Acknowledgments

This work was supported in part by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

Agarwal, S., Graepel, T., Herbrich, R., Har-Peled, S., & Roth, D. (2005). Generalization error bounds for the area under the roc curve. *Journal of Machine Learning Research*.

Amini, M.-R., & Gallinari, P. (2002). The use of unlabeled data to improve supervised learning for text summarization. *Proceedings of the 25th ACM SIGIR Conference* (pp. 105–112). Tampere, Finland.

Amini, M.-R., Usunier, N., & Gallinari, P. (2005). Automatic text summarization based on word-clusters and ranking algorithms. *Proceedings of the 27th European Conference on Information Retrieval (ECIR'05)*.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. *Proceedings of the 24th ACM SIGIR Conference* (pp. 276–284). New Orleans, Louisiana, United States.

Bartlett, P. L., & Mendelson, S. (2003). Rademacher and gaussian complexities: risk bounds and structural results. *J. Mach. Learn. Res.*, *3*, 463–482.

Caruana, R., & Niculescu-Mizil, A. (2004). Data mining in metric space: an empirical analysis of supervised learning performance criteria. *KDD '04: Proceedings of the 2004 ACM SIGKDD Conference* (pp. 69–78). Seattle, WA, USA.

Clémençon, S., Lugosi, G., & Vayatis, N. (2005). Ranking and scoring using empirical risk minimization. Preprint.

Cortes, C., & Mohri, M. (2004). Auc optimization vs. error rate minimization. In S. Thrun, L. Saul and B. Schölkopf (Eds.), *Advances in neural information processing systems 16*. Cambridge, MA: MIT Press.

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

Herschtal, A., & Raskutti, B. (2004). Optimising area under the roc curve using gradient descent. *ICML*.

Kupiec, J., Pedersen, J., & Chen, F. (1995). A trainable document summarizer. *Proceedings of the 18th ACM SIGIR Conference* (pp. 68–73). Seattle, Washington, United States.

McDiarmid, C. (1989). On the method of bounded differences. In *Surveys in combinatorics, 1989 (norwich, 1989)*, vol. 141 of *London Math. Soc. Lecture Note Ser.*, 148–188. Cambridge: Cambridge Univ. Press.

Meir, R., & Zhang, T. (2003). Generalization error bounds for bayesian mixture algorithms. *J. Mach. Learn. Res.*, *4*, 839–860.

Rakotomamonjy, A. (2004). Optimizing area under roc curve with svms. *ROCAI* (pp. 71–80).

Shawe-Taylor, J., & Critiani, N. (2004). *Kernel methods for pattern analysis*. Cambridge MA: MIT Press.

Vapnik, V. (1998). *Statistical learning theory*. New York: Springer-Verlag.

Vapnik, V. (2000). *The nature of statistical learning theory*. New York: Springer-Verlag.

Yan, L., Dodier, R., Mozer, M., & Wolniewicz, R. (2003). Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistics. *ICML-2003*.

Appendix A: proof of lemma 2

The proof of the lemma is an application of McDiarmid's Theorem (McDiarmid, 1989) which can be expressed as follows:

Theorem 7. ***((McDiarmid, 1989))*** *Let $X_1, ..., X_n$ be independent random variables taking values in a set A, and assume $f : A^n \to \mathbb{R}$ satisfies:*

$$\forall 1 \leq i \leq n \ :$$

$$\sup_{x_1,...,x_n,\hat{x}_i \in A} |f(x_1, ..., x_n) - f(x_1, ..., \hat{x}_i, x_{i+1}, ..., x_n)| \leq c_i$$

Then, for all $\epsilon > 0$:

$$P\{f(X_1, ..., X_n) - \mathbb{E}f(X_1, ..., X_n) \geq \epsilon\} \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2})$$

and:

$$P\{\mathbb{E}f(X_1, ..., X_n) - f(X_1, ..., X_n) \geq \epsilon\} \leq \exp(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2})$$

Now let us apply the theorem to the function

$$f : S \mapsto \sup_{Q \in \mathcal{Q}} (\mathbb{E}_{\mathcal{D}_1 \times \mathcal{D}_{-1}} Q - \hat{\mathbb{E}}_S Q) \tag{15}$$

where $S = (x_1, ..., x_n, x'_1, ...x'_m)$ is a sample of size $n+m$ drawn according to $\mathcal{D}_1^n \times \mathcal{D}_{-1}^m$ (that is, all x_i's and x'_j's are independent).

let $k \in \{1, ..., n\}$, $\hat{x}_k \in \mathcal{X}$ and S a sample of size $n+m$, and let $\hat{x}_i = x_i$ if $i \neq k$, and denote $\hat{S} = (\hat{x}_1, ..., \hat{x}_n, x'_1, ..., x'_m)$. Let furthermore Q be an element of $\mathcal{Q}$. Then, we have $\frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m Q(x_i, x'j) - \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m Q(\hat{x}_i, x'j) = \frac{1}{nm}\sum_{j=1}^m (Q(x_k, x'j) - Q(\hat{x}_k, x'_j))$. Since Q takes its values in $[0, 1]$, we have for all $j \mid Q(x_k, x'j) - Q(\hat{x}_k, x'_j) \mid \leq 1$, and, therefore $\mid \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m Q(x_i, x'j) - \frac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m Q(\hat{x}_i, x'j) \mid \leq \frac{1}{n}$, which can be written as $\mid \hat{\mathbb{E}}_S Q - \hat{\mathbb{E}}_{\hat{S}} Q \mid \leq \frac{1}{n}$.

Then, we have, for all Q in $\mathcal{Q}$ and all $\hat{x}_k$:

$$-\hat{\mathbb{E}}_{\hat{S}} Q - \tfrac{1}{n} \leq -\hat{\mathbb{E}}_S Q \leq -\hat{\mathbb{E}}_{\hat{S}} Q + \tfrac{1}{n}$$
$$\Rightarrow \mathbb{E}_{\mathcal{D}} Q - \hat{\mathbb{E}}_{\hat{S}} Q - \tfrac{1}{n} \leq \mathbb{E}_{\mathcal{D}} Q - \hat{\mathbb{E}}_S Q \leq \mathbb{E}_{\mathcal{D}} Q - \hat{\mathbb{E}}_{\hat{S}} Q + \tfrac{1}{n}$$

Taking the supremum over all Q, the last equation shows that, for all S, for all $\hat{x}_k$, we have:

$$\mid \sup_{Q \in \mathcal{Q}} (\mathbb{E}_{\mathcal{D}} Q - \hat{\mathbb{E}}_S Q) - \sup_{Q \in \mathcal{Q}} (\mathbb{E}_{\mathcal{D}} Q - \hat{\mathbb{E}}_{\hat{S}} Q) \mid \leq \frac{1}{n} \tag{16}$$

which can be expressed as:

$$\sup_{S, \hat{x}_k} \mid f(S) - f(x_1, ..., x_{k-1}, \hat{x}_k, x_{k+1}, ..., x_n, x'_1, ..., x'_m) \mid \leq \frac{1}{n} \tag{17}$$

where f is the function defined in equation 15 An analogous demonstration can be done to show that for $k \in \{1, ..., m\}$, to show that

$$\sup_{S, \hat{x}'_k} \mid f(S) - f(x_1, ..., x_n, x'_1, ..., x'_{k-1}, \hat{x}'_k, x'_{k+1}, ..., x'_m) \mid \leq \frac{1}{m} \tag{18}$$

Using McDiarmid's theorem with the results of equations 17 and 18, we can now say that:

$$\mathbb{P}_{\mathcal{D}_1^n \times \mathcal{D}_{-1}^m}(f(S) - \mathbb{E}f > \epsilon) \leq \exp(\frac{-2\epsilon^2}{n*(1/n)^2 + m*(1/m)^2})$$
$$\mathbb{P}_{\mathcal{D}_1^n \times \mathcal{D}_{-1}^m}(f(S) - \mathbb{E}f > \epsilon) \leq \exp(\frac{-2nm\epsilon^2}{n+m})$$

And solving for ϵ yields the result of lemma 2.

Appendix B: proof of lemma 3

In order for the proof to be readable, we first define some notations. n, m, $S = (\mathbf{x}, \mathbf{x}') = (x_1, ..., x_n, x'_1, x'_n)$, $\tilde{S} = (\tilde{\mathbf{x}}, \tilde{\mathbf{x}}') = (\tilde{x}_1, ..., \tilde{x}_n, \tilde{x}'_1, \tilde{x}'_n)$, σ, ν and $\mathcal{Q}$ have the same meaning as in lemma 3. We define the function $F(\sigma, \nu, S, \tilde{S}) = F(\sigma, \nu, \mathbf{x}, \mathbf{x}', \tilde{\mathbf{x}}, \tilde{\mathbf{x}}')$ to be equal to:

$$\sup_{Q \in \mathcal{Q}} \{\tfrac{1}{nm}\sum_{i=1}^n \sum_{j=1}^m \tfrac{\sigma_i + \nu_j}{2} Q(\tilde{x}_i, \tilde{x}'_j)$$
$$+ \tfrac{\sigma_i - \nu_j}{2} Q(\tilde{x}_i, x'_j) - \tfrac{\sigma_i - \nu_j}{2} Q(x_i, \tilde{x}'_j) - \tfrac{\sigma_i + \nu_j}{2} Q(x_i, x'_j))\} \tag{19}$$

We moreover denote by $\mathbf{X} = (X_i)_{i=1}^{2n} = (\mathbf{x}, \tilde{\mathbf{x}})$, that is $X_i = x_i$ if $1 \leq i \leq n$ and $X_i = \tilde{x}_{i-n}$ if $n+1 \leq i \leq 2n$, and $\mathbf{X}' = (X'_j)_{j=1}^{2m} = (\mathbf{x}', \tilde{\mathbf{x}}')$, such that $X'_j = x'_j$ if $1 \leq j \leq m$ and $X_j = \tilde{x}'_{j-m}$ if $m+1 \leq i \leq 2m$. When the context is clear, we will use the abuse of notation $F(\sigma, \nu, \mathbf{X}, \mathbf{X}')$ for $F(\sigma, \nu, (X_i)_{i=1}^n, (X'_j)_{j=1}^m, (X_i)_{i=n+1}^{2n}, (X'_j)_{j=m+1}^{2m})$. Moreover, if μ is a permutation of $\{1, ..., 2n\}$ and η a permutation of $\{1, ..., 2m\}$, we note $\mathbf{X}_\mu = (X_{\mu(i)})_{i=1}^{2n}$ and $\mathbf{X}'_\eta = (X'_{\eta(j)})_{j=1}^{2m}$. Finally, for a function $Q \in \mathcal{Q}$, $i \in \{1, ..., n\}$ and $j \in \{1, ..., m\}$, we define the term $F_Q^{\sigma,\nu,S,\tilde{S}}(i, j)$ (or $F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i, j)$ using an abuse of notation) to be equal to:

$$\tfrac{\sigma_i + \nu_j}{2} Q(\tilde{x}_i, \tilde{x}'_j) + \tfrac{\sigma_i - \nu_j}{2} Q(\tilde{x}_i, x'_j) - \tfrac{\sigma_i - \nu_j}{2} Q(x_i, \tilde{x}'_j)$$
$$- \tfrac{\sigma_i + \nu_j}{2} Q(x_i, x'_j))$$

or, alternatively (the two expressions are obviously equal):

$$\tfrac{\sigma_i + \nu_j}{2} Q(X_{n+i}, X'_{m+j}) + \tfrac{\sigma_i - \nu_j}{2} Q(X_{n+i}, X'_j)$$
$$- \tfrac{\sigma_i - \nu_j}{2} Q(X_i, X'_{m+j}) - \tfrac{\sigma_i + \nu_j}{2} Q(X_i, X'_j)) \tag{20}$$

such that we have:

$$F(\sigma, \nu, X, X') = \sup_{Q \in \mathcal{Q}} \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i, j) \tag{21}$$

As a base for our demonstration, we can claim, using standard probability arguments, that all the following

quantities are equal:

$$\mathbb{E}_{\mathbf{x},\mathbf{x}',\tilde{\mathbf{x}},\tilde{\mathbf{x}}'} F(\sigma,\nu,\mathbf{x},\mathbf{x}',\tilde{\mathbf{x}},\tilde{\mathbf{x}}') \quad (22)$$

$$\mathbb{E}_{S,\tilde{S}} F(\sigma,\nu,S,\tilde{S}) \quad (23)$$

$$\mathbb{E}_{\mathbf{X},\mathbf{X}'} F(\sigma,\nu,\mathbf{X},\mathbf{X}') \quad (24)$$

$$\mathbb{E}_{\mathbf{X},\mathbf{X}'} F(\sigma,\nu,\mathbf{X}_\mu,\mathbf{X}'_\eta) \quad (25)$$

We have to prove that $\mathbb{E}_{\sigma,\nu,S,\tilde{S}} F = \mathbb{E}_{S,\tilde{S}} \sup_{Q\in\mathcal{Q}} (\hat{\mathbb{E}}_{\tilde{S}} Q - \hat{\mathbb{E}}_S Q)$. In order to prove it, we can first notice that, noting $\sigma^{(0)} = (1)_{i=1}^n$ and $\nu^{(0)} = (1)_{j=1}^m$, we have, for all (i,j):

$$\begin{aligned} F_Q^{\sigma^{(0)},\nu^{(0)},\mathbf{X},\mathbf{X}'}(i,j) &= Q(X_{n+i}, X'_{n+j}) - Q(X_i, X'_j) \\ &= Q(\tilde{x}_i, \tilde{x}'_j) - Q(x_i, x'_j) \end{aligned}$$

and, as a consequence, using equality between the expressions of equation 23 and 24:

$$\begin{aligned} \mathbb{E}_{S,\tilde{S}} \sup_{Q\in\mathcal{Q}} (\hat{\mathbb{E}}_{\tilde{S}} Q - \hat{\mathbb{E}}_S Q) &= \mathbb{E}_{\mathbf{X},\mathbf{X}'} F(\sigma^{(0)},\nu^{(0)},\mathbf{X},\mathbf{X}') \\ &= \mathbb{E}_{S,\tilde{S}} F(\sigma^{(0)},\nu^{(0)},S,\tilde{S}) \quad (26) \end{aligned}$$

What remains to show is that

$$\mathbb{E}_{\sigma,\nu,\mathbf{X},\mathbf{X}'} F(\sigma,\nu,\mathbf{X},\mathbf{X}') = \mathbb{E}_{\mathbf{X},\mathbf{X}'} F(\sigma^{(0)},\nu^{(0)},\mathbf{X},\mathbf{X}') \quad (27)$$

In order to do that, we need the following notation: for $p \in \mathbb{N}$ and $k \in \{1,...,p\}$, let $\phi_{k,p}$ be the permutation of $\{1,...,2p\}$ defined by:

$$\phi_{k,p} : \left\{ \begin{array}{ll} \{1,...,2p\} \to \{1,...,2p\} & \\ k \mapsto k+p & \\ k+p \mapsto k & \mid k \in \{1,...,p\}\} \\ l \mapsto l \text{ if } l \notin \{k, p+k\} & \end{array} \right.$$

The demonstration of equation 27 can be done using the following lemma:

Lemma 8. *let $\sigma = (\sigma_i)_{i=1}^n$ and $\nu = (\nu_j)_{j=1}^m$ be two sequences of Rademacher variables. Now define μ and η, permutations of $\{1,...,2n\}$ and $\{1,...,2m\}$ respectively as:*

$$\mu = \bigcirc_{i=1}^n \phi_{i,n}^{\frac{1}{2}(1-\sigma_i)} \eta = \bigcirc_{j=1}^m \phi_{j,m}^{\frac{1}{2}(1-\nu_i)}$$

where ϕ^z denotes the power for the composition function, in particular, $\phi^1 = \phi$, $\phi^{-1} \circ \phi = Id$, $\phi^0 = Id$ where Id is the identity function. Then, we have, for all Q, $\mathbf{X}$, $\mathbf{X}'$, $i \in \{1,...,n\}$, $j \in \{1,...,m\}$:

$$F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i,j) = F_Q^{\sigma^{(0)},\nu^{(0)},\mathbf{X}_\mu,\mathbf{X}'_\eta}(i,j)$$

Proof. From equation 20, we have :

$$\begin{aligned} F_Q^{\sigma^{(0)},\nu^{(0)},\mathbf{X}_\mu,\mathbf{X}'_\eta}(i,j) = \quad & Q(X_{\mu(n+i)}, X'_{\eta(m+j)}) \quad (28) \\ & -Q(X_{\mu(i)}, X'_{\eta(j)}) \quad (29) \end{aligned}$$

It is to be noted that, since μ and η are the compositions of inversions $\phi_{k,p}$ which leave invariant all values different of k and $k+p$, we have $\mu(i) = i \wedge \mu(n+i) = n+i \Leftrightarrow \sigma_i = 1$, $\mu(i) = n+i \wedge \mu(i) = i \Leftrightarrow \sigma_i = -1$ [3], and similar equivalences for η. To prove the lemma, we therefore just need to consider the four cases which correspond to all the values the tuple (σ_i, ν_j) can take.

Case 1: $(\sigma_i, \nu_j) = (1,1)$. Then we have $F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i,j) = Q(X_{n+i}, X'_{m+j}) - Q(X_i, X'_j)$ from the definition of $F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}$ in equation 20. In this case, from the equivalences between the values of μ and η and σ_i and ν_i, we have $\mu(i) = i \wedge \mu(n+i) = n+i$ and $\eta(j) = j \wedge \eta(m+j) = m+j$ using the expression of equation 29:

$$\begin{aligned} F_Q^{\sigma^{(0)},\nu^{(0)},\mathbf{X}_\mu,\mathbf{X}'_\eta}(i,j) &= Q(X_{n+i}, X'_{m+j}) - Q(X_i, X'_j) \\ &= F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i,j) \end{aligned}$$

Case 2: $(\sigma_i, \nu_j) = (-1,1)$ From equation 20, we get $F_Q^{\sigma,\nu,\mathbf{X},\mathbf{X}'}(i,j) = Q(X_i, X'_{m+j}) - Q(X_{n+i}, X'_j)$. On the other hand, we have $\mu(i) = n+i$ and $\mu(n+i) = i$, while $\eta(j) = j$ and $\eta(m+j) = m+j$. Then, from equation 29, we have:

$$F_Q^{\sigma^{(0)},\nu^{(0)},\mathbf{X}_\mu,\mathbf{X}'_\eta}(i,j) = Q(X_i, X'_{m+j}) - Q(X_{n+i}, X'_j)$$

Hence showing the equality of the lemma in this case.

The two other cases are shown in the same way. Since only the four cases can appear, the equality of the lemma is proved. □

Back to the demonstration of lemma 3, using the result of lemma 8 and equation 21, we have:

$$F(\sigma,\nu,\mathbf{X},\mathbf{X}') = F(\sigma^{(0)},\nu^{(0)},\mathbf{X}_\mu,\mathbf{X}'_\eta)$$

Taking the expectation over $(\mathbf{X},\mathbf{X}')$ of this expression and using the equality between the expressions of equations 24 and 25, we have:

$$\mathbb{E}_{X,X'} F(\sigma,\nu,\mathbf{X},\mathbf{X}') = \mathbb{E}_{X,X'} F(\sigma^{(0)},\nu^{(0)},\mathbf{X},\mathbf{X}')$$

Taking the expectation over (σ,ν) and using the equality between the expressions of equations 24 and 23 leads to:

$$\mathbb{E}_{\sigma,\nu,S,\tilde{S}} F(\sigma,\nu,S,\tilde{S}) = \mathbb{E}_{S,\tilde{S}} F(\sigma^{(0)},\nu^{(0)},S,\tilde{S})$$

which, from the equality of equation 26 and the definition of F (equation 19) yields the result of the lemma.

[3] since there is an equivalence, $\mu(i)$ and $\mu(n+i)$ cannot take other values