Journal of Machine Learning Research 9 (2008) 2171-2185 Submitted 6/07; Revised 1/08; Published 10/08

A Moment Bound for Multi-hinge Classifiers

Bernadetta Tarigan TARIGAN@STAT.MATH.ETHZ.CH
Sara A. van de Geer GEER@STAT.MATH.ETHZ.CH
Seminar for Statistics
Swiss Federal Institute of Technology (ETH) Zurich
Leonhardstrasse 27, 8092 Zurich, Switzerland

Editor: Peter Bartlett

Abstract

The success of support vector machines in binary classification relies on the fact that hinge loss employed in the risk minimization targets the Bayes rule. Recent research explores some extensions of this large margin based method to the multicategory case. We show a moment bound for the so-called multi-hinge loss minimizers based on two kinds of complexity constraints: entropy with bracketing and empirical entropy. Obtaining such a result based on the latter is harder than finding one based on the former. We obtain fast rates of convergence that adapt to the unknown margin.

Keywords: multi-hinge classification, all-at-once, moment bound, fast rate, entropy

1. Introduction

We consider multicategory classification with equal cost. Let $Y \in \{1,\ldots,m\}$ denote one of the m possible categories, and let $X \in \mathbb{R}^d$ be a feature. We study the classification problem, where the goal is to predict Y given X with small error. Let $\{(X_i,Y_i)\}_{i=1}^n$ be an independent and identically distributed sample from (X,Y). In the binary case ($m=2$) a classifier $f : \mathbb{R}^d \to \mathbb{R}$ can be obtained by minimizing the empirical hinge loss

$$\frac{1}{n}\sum_{i=1}^{n}(1-Y_i f(X_i))_+ \tag{1}$$

over a given class of candidate classifiers $f \in \mathcal{F}$, where $(1-Yf(X))_+ := \max(0,1-Yf(X))$ with $Y \in \{\pm 1\}$. Hinge loss in combination with a reproducing kernel Hilbert space (RKHS) regularization penalty is called the support vector machine (SVM). See, for example, Evgeniou, Pontil, and Poggio (2000). In this paper, we examine the generalization of (1) to the multicategory case ($m > 2$). We refer to this classifier as the *multi-hinge*, although, instead of RKHS-regularization we will assume a given model class $\mathcal{F}$ satisfying a complexity constraint. We show a moment bound for the excess multi-hinge risk based on two kinds of complexity constraints: entropy with bracketing and empirical entropy. Obtaining such a result based on the latter is harder than finding one based on the former. We obtain fast rates of convergence that adapt to the unknown margin.

There are two strategies to generalize the binary SVM to the multicategory SVM. One strategy is by solving a series of binary problems; the other is by considering all of the categories at once. For the first strategy, some popular methods are the one-versus-rest method and the one-versus-one method. The one-versus-rest method constructs m binary SVM classifiers. The j-th classifier f_j is trained taking the examples from class j as positive and the examples from all other categories

as negative. A new example x is assigned to the category with the largest values of $f_j(x)$. The one-versus-one method constructs one binary SVM classifier for every pair of distinct categories, that is, all together $m(m-1)/2$ binary SVM classifiers are constructed. The classifier f_{ij} is trained taking the examples from category i as positive and the examples from category j as negative. For a new example x, if f_{ij} classifies x into category i then the vote for category i is increased by one. Otherwise the vote for category j is increased by one. After each of the $m(m-1)/2$ classifiers makes its vote, x is assigned to the category with the largest number of votes. See Duan and Keerthi (2005) and the references therein for an empirical study of the performance of these methods and its variants.

An all-at-once strategy for SVM loss has been proposed by some authors. For examples, see Vapnik (2000), Weston and Watkins (1999), Crammer and Singer (2000, 2001), and Guermeur (2002). Roughly speaking, the idea is similar to the one-versus-rest approach but all the m classifiers are obtained by solving one problem. (See Hsu and Lin, 2002, for details of the formulations.) Lee, Lin, and Wahba (2004) (see also Lee, 2002) show that the relationship of the formulations of the approaches above to the Bayes' rule is not clear from the literature and that they do not always implement the Bayes' rule. They propose a new approach that has good theoretical properties. That is, the defined loss is Bayes consistent and it provides a unifying framework for both equal and unequal misclassification costs.

We consider the equal misclassification cost where a correct classification costs 0 and an incorrect classification costs 1. The target function $f : \mathbb{R}^d \to \mathbb{R}^m$ is defined as an m-tuple of separating functions with zero-sum constraint $\sum_{j=1}^m f_j(x) = 0$, for any $x \in \mathbb{R}^d$. Hence, the classifier induced by $f(\cdot)$ is

$$g(\cdot) = \arg\max_{j=1,\ldots,m} f_j(\cdot)\ . \tag{2}$$

Analogous to the binary case, when applying RKHS-regularization, each component $f_j(x)$ is considered as an element of a RKHS $\overline{\mathcal{H}}_K = \{1\} + \mathcal{H}_K$, for all $j = 1,\ldots,m$. That is, $f_j(x)$ is expressed as $h_j(x) + b_j$ with $h_j \in \mathcal{H}_K$ and b_j some constant. To find $f(\cdot) = (f_1(\cdot),\ldots,f_m(\cdot)) \in \Pi_{j=1}^m \overline{\mathcal{H}}_K$ with the zero-sum constraint, the extension of SVM methodology is to minimize

$$\frac{1}{n}\sum_{i=1}^n \sum_{j=1, j\neq Y_i}^m (f_j(X_i) + \frac{1}{m-1})_+ \ + \ \frac{\lambda}{2}\sum_{j=1}^m ||h_j||^2_{\mathcal{H}_K}\ . \tag{3}$$

Based on (3), the multi-hinge loss is now defined as

$$l(Y, f(X)) := \sum_{j=1, j\neq Y}^m (f_j(X) + \frac{1}{m-1})_+\ . \tag{4}$$

The binary SVM loss (1) is a special case by taking $m = 2$. When $Y = 1$, $l(1, f(X)) = (f_2(X) + 1)_+ = (1 - f_1(X))_+$. Similarly, when $Y = -1$, $l(-1, f(X)) = (1 + f_1(X))_+$. Thus, (4) is identical with the binary SVM loss $(1 - Y f(X))_+$, where f_1 plays the same role as f.

Using a classifier g defined as in (2), a misclassification occurs whenever $g(X) \neq Y$. Let P be the unknown underlying measure of (X,Y). The prediction error of g is $P(g(X) \neq Y)$. Let $p_j(x)$ denote the conditional probability of category j given $x \in \mathbb{R}^d$, $j = 1,\ldots,m$. The prediction error is minimized by the Bayes classifier $g^* = \arg\max_{j=1,\ldots,m} p_j$, and the smallest prediction error is $P(g^*(X) \neq Y)$.

The theoretical multi-hinge risk is the expectation of the empirical multi-hinge loss with respect to the measure P and is denoted by

$$R(f) := \int l(y, f(x))\, dP(x,y)\ , \tag{5}$$

with $l(Y, f(X))$ defined as in (4). In this setting, Bayes' rule f^* is then an m-tuple separating functions with 1 in the kth coordinate and $-1/(m-1)$ elsewhere, whenever $k = \arg\max_{j=1,\ldots,m} p_j(x)$, $x \in \mathbb{R}^d$. Lemma 1 below shows that multi-hinge loss (4) is Bayes consistent. That is, f^* minimizes multi-hinge risk (5) over all possible classifiers. We write $R^* = R(f^*)$, the smallest possible multi-hinge risk. Lemma 1 is an extension of Bayes consistency of the binary SVM that has been shown by, for example, Lin (2002), Zhang (2004a) and Bartlett, Jordan, and McAuliffe (2006).

Lemma 1. *Bayes classifier f^* minimizes the multi-hinge risk $R(f)$.*

This lemma can be found in Lee, Lin, and Wahba (2004), Zhang (2004b,c), Tewari and Bartlett (2005) and Zou, Zhu, and Hastie (2006). We give a self-contained proof in Appendix for completeness. They establish the conditions needed to achieve the consistency for a general family of multicategory loss functions extended from various large margin binary classifiers. They also show that the SVM-type losses proposed by Weston and Watkins (1999) and Crammer and Singer (2001) are not Bayes consistent. Tewari and Bartlett (2005) and Zhang (2004b,c) also show that the convergence to zero (in probability) of the excess multi-hinge risk $R(f) - R^*$ implies the convergence to zero with the same rate (in probability) of the excess prediction error $P(g(f(X)) \neq Y) - P(g(f^*(X)) \neq Y)$.

The RKHS-regularization (3) has attracted some interest. For example, Lee and Cui (2006) study an algorithm of fitting the entire regularization path and Wang and Shen (2007) study the use of l_1 penalty in place of the l_2 penalty. In this paper, we will not study the RKHS-regularization but we take the minimization of the empirical multi-hinge loss over a given class of candidate classifiers $\mathcal{F}$ satisfying a complexity constraint. That is, we do not invoke a penalization technique.

Let $\mathcal{F}$ be a model class of candidate classifiers. For $j = 1, \ldots, m$, we assume that each f_j is a member of the same class $\mathcal{F}_o = \{h : \mathbb{R}^d \to \mathbb{R}, h \in L_2(Q)\}$, with Q the unknown marginal distribution of X. That is,

$$\mathcal{F} = \{f = (f_1, \ldots, f_m) : \sum_{j=1}^{m} f_j = 0,\ f_j \in \mathcal{F}_o\}\ . \tag{6}$$

Let P_n be the empirical distribution of (X,Y) based on the observations $\{(X_i, Y_i)\}_{i=1}^n$ and Q_n the corresponding empirical distribution of X based on $X_1, \ldots, X_n$. We endow $\mathcal{F}$ with the following squared semi-metrics

$$\begin{aligned} \|f - \tilde{f}\|_{2,Q}^2 &:= \sum_{j=1}^{m} \int |f_j - \tilde{f}_j|^2\, dQ\ , \text{ and} \\ \|f - \tilde{f}\|_{2,Q_n}^2 &:= \sum_{j=1}^{m} \frac{1}{n} \sum_{i=1}^{n} |f_j(X_i) - \tilde{f}_j(X_i)|^2\ , \end{aligned}$$

for all $f, \tilde{f} \in \mathcal{F}$. We impose a complexity constraint on the class $\mathcal{F}_o$ in term of either the entropy with bracketing or the empirical entropy. Below we give the definitions of the entropies.

Definition of entropy. *Let $\mathcal{G}$ be a subset of a metric space (Λ, d). Let*

$$H(\varepsilon, \mathcal{G}, d) := \log N(\varepsilon, \mathcal{G}, d) \text{ , for all } \varepsilon > 0 \text{ ,}$$

where $N(\varepsilon, \mathcal{G}, d)$ is the smallest value of N for which there exist functions $g_1, \ldots, g_N$ in $\mathcal{G}$, such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$, such that

$$d(g, g_j) \leq \varepsilon \text{ .}$$

Then $N(\varepsilon, \mathcal{G}, d)$ is called the ε-covering number *of $\mathcal{G}$ and $H(\varepsilon, \mathcal{G}, d)$ is called the* ε-entropy *of $\mathcal{G}$ (for the d-metric).*

Definition of entropy with bracketing. *Let $\mathcal{G}$ be a subset of a metric space (Λ, d) of real-valued functions. Let*

$$H_B(\varepsilon, \mathcal{G}, d) := \log N_B(\varepsilon, \mathcal{G}, d) \text{ , for all } \varepsilon > 0 \text{ ,}$$

where $N_B(\varepsilon, \mathcal{G}, d)$ is the smallest value of N for which there exist pairs of functions $\{[g_1^L, g_1^U], \ldots, [g_N^L, g_N^U]\}$ such that $d(g_j^L, g_j^U) \leq \varepsilon$ for all $j = 1, \ldots, N$, and such that for each $g \in \mathcal{G}$, there is a $j = j(g) \in \{1, \ldots, N\}$ such that

$$g_j^L \leq g \leq g_j^U \text{ .}$$

Then $N_B(\varepsilon, \mathcal{G}, d)$ is called the ε-covering number with bracketing *of $\mathcal{G}$ and $H_B(\varepsilon, \mathcal{G}, d)$ is called the* ε-entropy with bracketing *of $\mathcal{G}$ (for the d-metric).*

Let $H_B(\varepsilon, \mathcal{F}_o, L_2(Q))$ and $H(\varepsilon, \mathcal{F}_o, L_2(Q_n))$ denote the ε-entropy with bracketing and the empirical ε-entropy of the class $\mathcal{F}_o$, respectively. The complexity of a model class can be summarized in a complexity parameter $\rho \in (0, 1)$. Let A be some positive constant. We consider classes $\mathcal{F}_o$ satisfying one of the following complexity constraints:

$$\begin{aligned} H_B(\varepsilon, \mathcal{F}_o, L_2(Q)) &\leq A\varepsilon^{-2\rho} \text{ , for all } \varepsilon > 0 \text{ , or} \\ H(\varepsilon, \mathcal{F}_o, L_2(Q_n)) &\leq A\varepsilon^{-2\rho} \text{ , for all } \varepsilon > 0 \text{ , a.s. for all } n \geq 1 \text{ .} \end{aligned}$$

It is straightforward to show that for all $\varepsilon > 0$:

$$\begin{aligned} H_B(\varepsilon, \mathcal{F}, \|\cdot\|_{2,Q}) &\leq (m-1)\, H_B(\varepsilon(m-1)^{-1/2},\ \mathcal{F}_o,\ L_2(Q)) \text{ ,} \\ H(\varepsilon, \mathcal{F}, \|\cdot\|_{2,Q_n}) &\leq (m-1)\, H(\varepsilon(2(m-1))^{-1/2},\ \mathcal{F}_o,\ L_2(Q_n)) \text{ .} \end{aligned}$$

We define the minimizer of the empirical multi-hinge loss (without penalty)

$$\hat{f}_n := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1, j \neq Y_i}^{m} (f_j(X_i) + \frac{1}{m-1})_+ \text{ ,} \tag{7}$$

where the model class $\mathcal{F}$ defined as in (6) satisfies either an entropy with bracketing constraint or an empirical entropy constraint described above.

Besides the model class complexity, the rate of convergence also depends on the so-called margin condition (see Condition A below) that quantifies the identifiability of the Bayes rule and is summarized in a margin parameter (or noise level) $\kappa \geq 1$. In Tarigan and van de Geer (2006), a

probability inequality has been obtained for l_1-penalized excess hinge risk in the binary case that adapts to the unknown parameters. In this paper, we show a moment bound for the excess multi-hinge risk $R(\hat{f}_n) - R^*$ of $\hat{f}_n$ over the model class $\mathcal{F}$ with rate of convergence $n^{-\kappa/(2\kappa-1+\rho)}$, which is faster than $n^{-1/2}$.

In Section 2 we present our main result based on the margin and complexity conditions. The proof of the main result is given in Section 3, together with our supporting lemmas. For the sake of completeness and to avoid distraction, we place the proof of some supporting lemmas in the Appendix.

2. A Moment Bound for Multi-hinge Classifiers

We first state the margin and the complexity conditions.

Condition A (Margin condition)**.** *There exist constants $\sigma > 0$ and $\kappa \geq 1$ such that for all $f \in \mathcal{F}$,*

$$R(f) - R^* \;\geq\; \frac{1}{\sigma^\kappa}\Big(\sum_{j=1}^{m} \int |f_j - f_j^*| \, dQ\Big)^\kappa .$$

Condition B1 (Complexity constraint under ε-entropy with bracketing)**.** *Let $0 < \rho < 1$ and let A be a positive constant. The ε-entropy with bracketing satisfies the inequality*

$$H_B(\varepsilon, \mathcal{F}_o, L_2(Q)) \leq A\varepsilon^{-2\rho} \;, \;\; \textit{for all } \varepsilon > 0 \, .$$

Condition B2 (Complexity constraint under empirical ε-entropy)**.** *Let $0 < \rho < 1$ and let A be a positive constant. The empirical ε-entropy, almost surely for all $n \geq 1$, satisfies the inequality*

$$H(\varepsilon, \mathcal{F}_o, L_2(Q_n)) \leq A\varepsilon^{-2\rho} \;, \;\; \textit{for all } \varepsilon > 0 \, .$$

Now we come to the main result.

Theorem 2. *Assume Condition A is met and that $|f_j - f_j^*| \leq M$ for all $j = 1, \ldots, m$, and all $f = (f_1, \ldots, f_m) \in \mathcal{F}$. Let $\hat{f}_n$ be the multi-hinge loss minimizer defined in (7). Suppose that either Condition B1 or Condition B2 holds. Then for small values of $\delta > 0$,*

$$\mathbb{E}[R(\hat{f}_n) - R^*] \leq \frac{1+\delta}{1-\delta} \, \inf\Big\{ R(f) - R^* + C_0 \, n^{-\frac{\kappa}{2\kappa-1+\rho}} \;:\; f \in \mathcal{F} \Big\}$$

with C_0 some constant depending only on m, M, κ, σ, A and ρ.

Condition A follows from the condition on the behaviour of the conditional probabilities p_j. We formulate this in Condition AA below. We require that, for a fixed $x \in \mathbb{R}^d$, there is no pair of categories having the same conditional probabilities each of which stays away from 1. Originally

the terminology "margin condition" comes from the binary case of the prediction error considered in the work of Mammen and Tsybakov (1999) and Tsybakov (2004), where the behaviour of p_1, the conditional probability of category 1, is restricted near $\{x : p_1(x) = 1/2\}$. The "margin" set $\{x : p_1(x) = 1/2\}$ identifies the Bayes predictor which assigns a new x to class 1 if $p_1(x) > 1/2$ and class 2 otherwise. The margin condition is also called the *condition on the noise level*, and it is summarized in a margin parameter κ. Boucheron, Bousquet, and Lugosi (2005, Section 5.2) discuss the noise condition and its equivalent variants, corresponding to the fast rates of convergence, in the binary case. Thus, Condition AA is a natural extension for the multicategory case wrt. hinge loss. Lemma 3 below gives the connection between Condition A and Condition AA. We provide the proof in the Appendix. For $x \in \mathcal{X}$, let $p_k(x) = \max_{j\in\{1,\ldots,m\}} p_j(x)$ and define

$$\tau(x) := \min_{j\neq k}\{|p_j(x) - p_k(x)|, 1 - p_k(x)\}\ , \tag{8}$$

where j and k take values in $\{1,2,\ldots,m\}$.

Condition AA. *Let τ be defined in (8). There exist constants $C \geq 1$ and $\gamma \geq 0$ such that $\forall z > 0$,*

$$Q(\{\tau \leq z\}) \leq (Cz)^{1/\gamma}\ .$$

[Here we use the convention $(Cz)^{1/\gamma} = \mathbb{1}\{z \geq 1/C\}$ for $\gamma = 0$.]

Lemma 3. *Suppose Condition AA is met. Then for all $f \in \mathcal{F}$ with $|f_j - f_j^*| \leq M$ for all $j = 1,\ldots,m$,*

$$R(f) - R^* \geq \frac{1}{\sigma_M}\Big(\sum_{j=1}^{m}\int |f_j - f_j^*|\, dQ\Big)^{1+\gamma}\ ,$$

where $\sigma_M = C\,(mM(1/\gamma+1))^{\gamma}(1+\gamma)$. That is, Condition A holds with $\sigma = (\sigma_M)^{1/\kappa}$ and $\kappa = 1+\gamma$.

Remark. In the definition of τ we have the extra piece $1 - p_k$. It is needed for technical reason. It forces that nowhere in the input space one class can clearly dominate. We refer to the work of Bartlett and Wegkamp (2006, Section 4) and Tarigan and van de Geer (2006, Section 3.3.1) for some ideas how to get around this difficulty.

The complexity constraints B1 and B2 cover some interesting classes, including Vapnik-Chervonenkis (VC) subgraph classes and VC convex hull classes. See, for example, van der Vaart and Wellner (1996, Section 2.7), van de Geer (2000, Sections 2.4, 3.7, 7.4, 10.1 and 10.3) and Song and Wellner (2002). In the situation when the approximation error $\inf_{f\in\mathcal{F}} R(f) - R^*$ is zero (the model class $\mathcal{F}$ contains the Bayes classifier), Steinwart and Scovel (2005) obtain the same rate of convergence for the excess hinge risk under the margin condition A and the complexity condition B2. They consider the RKHS-regularization setting for the binary case instead.

We do not explore the behaviour of the approximation error $\inf_{f\in\mathcal{F}} R(f) - R^*$. This problem is still open and very hard to solve even in the binary case.

3. Proof of Theorem 2

Let $f^o := \arg\min_{f\in\mathcal{F}} R(f)$, the minimizer of the theoretical risk in the model class $\mathcal{F}$. As shorthand notation we write for the loss $l_f = l_f(X,Y) = l(Y, f(X))$. We also write $\nu_n(l_f) = \sqrt{n}\,(R_n(f) - R(f))$.

Since $R_n(\hat{f}_n) - R_n(f) \leq 0$ for all $f \in \mathcal{F}$, we have

$$\begin{aligned} R(\hat{f}_n) - R^* &\leq -[R_n(\hat{f}_n) - R(\hat{f}_n)] + [R_n(f^o) - R(f^o)] + R(f^o) - R^* \\ &\leq |\nu_n(l_{\hat{f}_n}) - \nu_n(l_{f^o})|/\sqrt{n} + R(f^o) - R^* \ . \end{aligned} \tag{9}$$

We call inequality (9) a basic-inequality, following van de Geer (2000). This upper bound enables us to work with the increments of the empirical process $\{\nu_n(l_f) - \nu_n(l_{f^o}) \ : \ l_f \in \mathcal{L}\}$ indexed by the multi-hinge loss $l_f \in \mathcal{L}$, where $\mathcal{L} = \{l_f : f \in \mathcal{F}\}$.

The procedure of the proof is based on the proof of Lemma 2.1 in del Barrio et al. (2007), page 206. We write

$$Z_n(l_f) := \frac{|\nu_n(l_f) - \nu_n(l_{f^o})|}{\left(\|l_f - l_{f^o}\|_{2,P} \vee n^{-\frac{1}{2+2\rho}}\right)^{1-\rho}} \ , \ l_f \in \mathcal{L} \ ,$$

where $(a \vee b) := \max\{a,b\}$, $\|l_f\|_{2,P}^2 := \int l_f^2(x,y)\, dP(x,y)$ and ρ is from either Condition B1 or B2. For short hand of notation, we also write $Z_n = Z_n(l_{\hat{f}_n})$. Then

$$R(\hat{f}_n) - R^* \leq (Z_n/\sqrt{n}) \ (\ \|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho} \vee n^{-\frac{1-\rho}{2+2\rho}}) \ + \ R(f^o) - R^* \ . \tag{10}$$

Applying the triangular inequality and Lemma 4 below gives

$$\|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho} \leq (m-1)^{(1-\rho)/2} \ \left(\|\hat{f}_n - f^*\|_{2,Q}^{1-\rho} + \|f^o - f^*\|_{2,Q}^{1-\rho}\right) .$$

Observe that for any $f \in \mathcal{F}$ with $|f_j - f_j^*| \leq M$, and for all j, Condition A gives $\|f - f^*\|_{2,Q}^2 \leq M\sigma \ (R(f) - R^*)^{1/\kappa}$. Thus,

$$\|l_{\hat{f}_n} - l_{f^o}\|_{2,P}^{1-\rho} \leq C_1 \left\{[R(\hat{f}_n) - R^*]^{(1-\rho)/2\kappa} + [R(f^o) - R^*]^{(1-\rho)/2\kappa}\right\} ,$$

with $C_1 = ((m-1)M\sigma)^{(1-\rho)/2}$. Denote by $\mathcal{R}$ the right hand side of the above inequality. Hence, from (10) we have

$$R(\hat{f}_n) - R^* \leq (Z_n/\sqrt{n}) \ (\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) + R(f^o) - R^* \ .$$

We consider first the case $(\mathcal{R} \vee n^{-\frac{1-\rho}{2+2\rho}}) = \mathcal{R}$. That is,

$$R(\hat{f}_n) - R^* \leq \frac{Z_n}{\sqrt{n}} C_1 \left\{[R(\hat{f}_n) - R^*]^{(1-\rho)/2\kappa} + [R(f^o) - R(f^*)]^{(1-\rho)/2\kappa}\right\} + R(f^o) - R^* \ .$$

Two applications of Lemma 5 below yield for all $0 < \delta < 1$,

$$\begin{aligned} R(\hat{f}_n) - R^* &\leq \delta(R(\hat{f}_n) - R^*) + (1+\delta)(R(f^o) - R^*) + 2\left(C_1 \ Z_n/\sqrt{n}\right)^{\frac{2\kappa}{2\kappa-1+\rho}} \ \delta^{-\frac{1-\rho}{2\kappa-1+\rho}} \\ &\leq \delta(R(\hat{f}_n) - R^*) \ + (1+\delta)\left(R(f^o) - R^* + C_2 \ Z_n^r \ n^{-\frac{\kappa}{2\kappa-1+\rho}}\right) , \end{aligned}$$

with $C_2 = 2 \ C_1^r \ \delta^{-\frac{1-\rho}{2\kappa-1+\rho}}$ and $r = 2\kappa/(2\kappa-1+\rho)$. Now it is left to show that $\mathbb{E}[Z_n^r]$ is bounded, say by some constant C_3. Then, $C_0 = C_2 C_3$ in Theorem 2.

To show that $\mathbb{E}[Z_n^r]$ is bounded, we use an exponential tail probability of the supremum of the weighted empirical process

$$\{Z_n(l_f) \ : \ l_f \in \mathcal{L}\} \ . \tag{11}$$

We recall that $H_B(\varepsilon, \mathcal{F}, \|\cdot\|_{2,Q}) \leq (m-1)H_B(\varepsilon(m-1)^{-1/2}, \mathcal{F}_o, L_2(Q))$. A key observation is that

$$H_B(\varepsilon, \mathcal{L}, L_2(P)) \leq (m-1)\, H_B(\varepsilon(m-1)^{-1/2}, \mathcal{F}, \|\cdot\|_{2,Q})\,,$$

by Lemma 4. It gives an upper bound for the ε-entropy with bracketing of the model class $\mathcal{L}$: $H_B(\varepsilon, \mathcal{L}, L_2(P)) \leq A_o\varepsilon^{-2\rho}$, for all $\varepsilon > 0$, with $A_o = A(m-1)^{2+2\rho}$. Under Condition B1, an application of Lemma 5.14 in van de Geer (2000), presented below in Lemma 6, gives the desired exponential tail probability. Hence, for some positive constant c,

$$\begin{aligned}\mathbb{E}[Z_n^r] &= \int_0^c \mathbb{P}(Z_n \geq t^{1/r})\, dt \,+\, \int_c^\infty \mathbb{P}(Z_n \geq t^{1/r})\, dt \\ &\leq c \,+\, \int_0^\infty c\, \exp(-\frac{t^{1/r}}{c^2})\, dt \,= c + rc^{2r+1}\Gamma(r)\,.\end{aligned}$$

For the case $\mathcal{R} \leq n^{-(1-\rho)/(2+2\rho)}$, we have

$$R(\hat{f}_n) - R^* \leq Z_n\, n^{-1/(1+\rho)} \,+\, R(f^o) - R^*\,.$$

We conclude by noting that $n^{-1/(1+\rho)} \leq n^{-\kappa/(2\kappa-1+\rho)}$, where $\kappa \geq 1$ and $0 < \rho < 1$.

Now we consider the case where Condition B2 holds instead of B1. By virtue of the proof above, we need only to verify an exponential probability of the supremum of the process (11) under Condition B2 instead of B1. This is done by employing Lemmas 7–9 below. Again, a key observation is that Lemma 4 and Condition B2 give us $H(\varepsilon, \mathcal{L}, L_2(P_n)) \leq A(m-1)^{2+2\rho}\varepsilon^{-2\rho}$. ■

Lemma 4 gives an upper bound of the squared $L_2(P)$-metric of the excess loss in terms of $\|\cdot\|_{2,Q}$-metric.

Lemma 4. $\mathbb{E}[(l_f(X,Y) - l_{f^*}(X,Y))^2] \leq (m-1)\sum_{j=1}^m \int |f_j - f_j^*|^2\, dQ\,.$

Proof. We write $\Delta(f, f^*) = \mathbb{E}_{Y|X}[(l_f(X,Y) - l_{f^*}(X,Y))^2 | X = x]$ and recall that $p_j(x) = P(Y = j | X = x)$, for all $j = 1, \ldots, m$. We fix an arbitrary $x \in \mathbb{R}^d$. Definition of the loss gives

$$\begin{aligned}\Delta(f, f^*) &= \sum_{j=1}^m p_j \Big(\sum_{i\neq j}(f_i + \frac{1}{m-1})_+ - (f_i^* + \frac{1}{m-1})_+\Big)^2 \\ &= \sum_{j=1}^m p_j \Big(\sum_{i\in I^+(j)}(f_i - f_i^*) + \sum_{i\in I^-(j)}(-\frac{1}{m-1} - f_i^*)\Big)^2\,,\end{aligned}$$

where $I^+(j) = \{i \neq j \;:\; f_i \geq -1/(m-1),\ i = 1, \ldots, m\}$ and $I^-(j) = \{i \neq j \;:\; f_i < -1/(m-1),\ i = 1, \ldots, m\}$. Use the facts that $(\sum_{i=1}^n a_i)^2 \leq n\sum_{i=1}^n a_i^2$ for all $n \in \mathbb{N}$ and $a_i \in \mathbb{R}$, and that $\max\{|I^+(j)|, |I^-(j)|\} \leq m-1$, to obtain

$$\Delta(f, f^*) \leq (m-1)\sum_{j=1}^m p_j \Big(\sum_{i\in I^+(j)}(f_i - f_i^*)^2 \,+\, \sum_{i\in I^-(j)}(-\frac{1}{m-1} - f_i^*)^2\Big)\,.$$

Clearly, $|-1/(m-1) - f_i^*| \leq |f_i - f_i^*|$ for all $i \in I^-(j)$. Hence,

$$\Delta(f, f^*) \leq (m-1)\sum_{j=1}^m p_j\, (\sum_{i\neq j}|f_i - f_i^*|^2) = (m-1)\sum_{j=1}^m (1-p_j)|f_j - f_j^*|^2\,,$$

where the last equality is obtained using $\sum_{j=1}^{m} p_j = 1$. We conclude the proof by bounding $1 - p_j$ with 1 for all j and integrating over all $x \in \mathbb{R}^d$ wrt. the marginal distribution Q. ■

The technical lemma below is an immediate consequence of Young's inequality (see, for example, Hardy, Littlewood, and Pólya, 1988, Chapter 8.3), using some straightforward bounds to simplify the expressions.

Lemma 5 (Technical Lemma)**.** *For all positive* ν, t, δ *and* $\kappa > \beta$*:*

$$\nu t^{\beta/\kappa} \;\leq\; \delta t \;+\; \nu^{\frac{\kappa}{\kappa-\beta}}\, \delta^{\frac{-\beta}{\kappa-\beta}} \;.$$

To ease the exposition, throughout Lemma 6 and Lemma 7 we write $\|\cdot\| = \|\cdot\|_{2,Q}$ and $\|\cdot\|_n = \|\cdot\|_{2,Q_n}$ for the $L_2(Q)$-norm and the $L_2(Q_n)$-norm, respectively.

Lemma 6 (van de Geer, 2000, Lemma 5.14)**.** *For a probability measure Q, let $\mathcal{H}$ be a class of uniformly bounded functions h in $L_2(Q)$, say $\sup_{h\in\mathcal{H}} |h - h^o|_\infty < 1$, where h^o is a fixed but arbitrary function in $\mathcal{H}$. Suppose that*

$$H_B(\varepsilon, \mathcal{H}, L_2(Q)) \leq A_o \varepsilon^{-2\rho} \;,\; \textit{for all } \varepsilon > 0 \;,$$

with $0 < \rho < 1$ and $A_o > 0$. Then for some positive constants c and n_o depending only on ρ and A_o,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h - h^o\| \vee n^{-\frac{1}{2+2\rho}} \Big)^{1-\rho}} \geq t \Big) \leq c\, \exp(-t/c^2) \;,$$

for all $t > c$ and $n > n_o$.

Lemma 7. *For a probability measure Q on $(\mathcal{Z}, \mathcal{A})$, let $\mathcal{H}$ be a class of uniformly bounded functions h in $L_2(Q)$, say $\sup_{h\in\mathcal{H}} |h - h^o|_\infty < 1$, where h^o is a fixed but arbitrary element in $\mathcal{H}$. Suppose that*

$$H(\varepsilon, \mathcal{H}, L_2(Q_n)) \leq A_o \varepsilon^{-2\rho} \;,\; \textit{for all } \varepsilon > 0 \;,$$

with $0 < \rho < 1$ and $A_o > 0$. Then for some positive constants c and n_o depending on ρ and A_o,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h - h^o\| \vee n^{-\frac{1}{2+2\rho}} \Big)^{1-\rho}} \geq t \Big) \leq c\, \exp(-t/c^2) \;,$$

for all $t > c$ and $n > n_o$.

Proof. For $n \geq (t^2/8)^{1+\rho/(1-\rho)}$, Chebyshev's inequality and a symmetrization technique (see, for example, van de Geer, 2000, page 32) give

$$\mathbb{P}\left(\sup_{h\in\mathcal{H}} \frac{|\nu_n(h) - \nu_n(h^o)|}{\Big(\|h - h^o\| \vee n^{-1/(2+2\rho)} \Big)^{1-\rho}} \geq t \right)$$

$$\leq 4\mathbb{P}\left(\sup_{h\in\mathcal{H}}\frac{|\nu_n^\varepsilon(h)-\nu_n^\varepsilon(h^o)|}{\left(\|h-h^o\|_n\vee n^{-1/(2+2\rho)}\right)^{1-\rho}}\geq\sqrt{t}/4\right) \tag{12}$$

$$+\,4\mathbb{P}\left(\sup_{h\in\mathcal{H}}\frac{\|h-h^o\|_n^{1-\rho}}{\left(\|h-h^o\|\vee n^{-1/(2+2\rho)}\right)^{1-\rho}}\geq\sqrt{t}/4\right), \tag{13}$$

where $\nu_n^\varepsilon(h)$ is the symmetrized version of the $\nu_n(h)$. That is, $\nu_n^\varepsilon(h)=(1/\sqrt{n})\sum_{i=1}^n\varepsilon_i h(Z_i)$, where $\{\varepsilon_i\}_{i=1}^n$ are independent random variables, independent of $\{Z_i\}_{i=1}^n$, with $\mathbb{P}(\varepsilon_i=1)=\mathbb{P}(\varepsilon_i=-1)=1/2$ for all $i=1,\ldots,n$.

To handle (12), we divide the class $\mathcal{H}$ into two disjoint classes where the empirical distance $\|h-h^o\|_n$ is smaller or larger than $n^{-1/(2+2\rho)}$. Write $\mathcal{H}_n=\{h\in\mathcal{H}:\ \|h-h^o\|_n\leq n^{-1/(2+2\rho)}\}$. By Lemma 5.1 in van de Geer (2000), stated below in Lemma 8, for some positive constant c_1,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}_n}\frac{|\nu_n^\varepsilon(h)-\nu_n^\varepsilon(h^o)|}{n^{-(1-\rho)/(2+2\rho)}}\geq\sqrt{t}/4\Big)\leq c_1\exp\Big(-\frac{t\,n^{1/(1+\rho)}}{64\,c_1^2}\Big)\,.$$

Let $J=\min\{j>1:2^{-j}<n^{-1/(2+2\rho)}\}$. We apply the peeling device on the set $\{h\in\mathcal{H}:2^{-j}\leq\|h-h^o\|_n\leq 2^{-j+1},\ j=1,\ldots,J\}$ to obtain that, for all $t>1$,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}_n^c}\frac{|\nu_n^\varepsilon(h)-\nu_n^\varepsilon(h^o)|}{\|h-h^o\|_n^{1-\rho}}\geq\sqrt{t}/4\ \Big|\ Z_1,\ldots,Z_n\Big)$$

$$\leq\sum_{j=1}^J\mathbb{P}\Big(\sup_{\substack{h\in\mathcal{H}\\ \|h-h^o\|_n\leq 2^{-j+1}}}|\nu_n^\varepsilon(h)-\nu_n^\varepsilon(h^o)|\geq\frac{\sqrt{t}}{4}\,2^{-j(1-\rho)}\mid Z_1,\ldots,Z_n\Big)$$

$$\leq\sum_{j=1}^J c_2\exp(-\frac{t\,2^{2\rho j}}{216\,c_2^2})\ \leq c\ \exp(-t/c^2)\,.$$

To handle (13), we use a modification of Lemma 5.6 in van de Geer (2000), stated below in Lemma 9, where we take t such that $(\sqrt{t}/4)^{1/(1-\rho)}\geq 14u$. ■

Lemma 8 (van de Geer, 2000, Lemma 5.1)**.** *Let $Z_1,\ldots,Z_n,\ldots$ be i.i.d. with distribution Q on $(\mathcal{Z},\mathcal{A})$. Let $\{\varepsilon_i\}_{i=1}^n$ be independent random variables, independent of $\{Z_i\}_{i=1}^n$, with $\mathbb{P}(\varepsilon_i=1)=\mathbb{P}(\varepsilon_i=-1)=1/2$ for all $i=1,\ldots,n$. Let $\mathcal{H}\subset L_2(Q)$ be a class of functions on $\mathcal{Z}$. Write $\nu_n^\varepsilon(h):=\frac{1}{\sqrt{n}}\sum_{i=1}^n\varepsilon_i h(Z_i)$, with $h\in\mathcal{H}$. Let*

$$\mathcal{H}(\delta):=\{h\in\mathcal{H}:\|h-h^o\|_{2,Q}\leq\delta\}\,,\quad\hat{\delta}_n:=\sup_{h\in\mathcal{H}(\delta)}\|h-h^o\|_{2,Q_n}\,,$$

where h^o is a fixed but arbitrary function in $\mathcal{H}$ and Q_n is the corresponding empirical distribution of Z based on $\{Z_i\}_{i=1}^n$. For $a\geq 8C\left(\int_{a/(32\sqrt{n})}^{\hat{\delta}_n}H^{1/2}(u,\mathcal{H},Q_n)du\ \vee\ \hat{\delta}_n\right)$, where C is some positive constant, we have

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}(\delta)}|\nu_n^\varepsilon(h)-\nu_n^\varepsilon(h^o)|\geq\frac{a}{4}\Big|Z_1,\ldots,Z_n\Big)\leq C\exp(-\frac{a^2}{64C^2\hat{\delta}_n^2})\,.$$

The following lemma is a modification of Lemma 5.6 in van de Geer (2000).

Lemma 9. *For a probability measure S on $(\mathcal{Z},\mathcal{A})$, let $\mathcal{H}$ be a class of uniformly bounded functions independent of n with $\sup_{h\in\mathcal{H}}|h|_\infty \le 1$. Suppose that almost surely for all $n \ge 1$,*

$$H(\varepsilon,\mathcal{H},L_2(S_n)) \le A_o\varepsilon^{-2\rho}, \text{ for all } \varepsilon > 0 ,$$

with $0<\rho<1$ and $A_o>0$. Then, for all n,

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}} \frac{\|h\|_{2,S_n}}{\|h\|_{2,S}\vee n^{-\frac{1}{2+2\rho}}} \ge 14u\Big) \le 4\,\exp(-u^2\, n^{\frac{\rho}{1+\rho}}) ,$$

for all $u \ge 1$.

Proof. Let $\{\delta_n\}$ be a sequence with $\delta_n\to 0$, $n\delta_n^2\to\infty$, $n\delta_n^2 \ge 2A_oH(\delta_n)$ for all n with $H(\delta_n)=\delta_n^{-2\rho}$. We apply the randomization device in Pollard (1984, page 32), as follows. Let $Z_{n+1},\dots,Z_{2n}$ be an independent copy of $Z_1,\dots,Z_n$. Let $\omega_1,\dots,\omega_n$ be independent random variables, independent of $Z_1,\dots,Z_{2n}$, with $\mathbb{P}(\omega_i=1)=\mathbb{P}(\omega_i=0)=1/2$ for all $i=1,\dots,n$. Set $Z_i{}' = Z_{2i-1+\omega_i}$ and $Z_i{}'' = Z_{2i-\omega_i}$, $i=1,\dots,n$, and $S_n{}' = (1/n)\sum_{i=1}^n \delta_{Z_i'}$, $S_n{}'' = (1/n)\sum_{i=1}^n \delta_{Z_i''}$, and $\bar{S}_{2n} = (S_n{}' + S_n{}'')/2$. Since the class is uniformly bounded by 1, an application of Chebyshev's inequality gives that for each h in $\mathcal{H}$,

$$\mathbb{P}\Big(\frac{\|h\|_{2,S_n}}{\|h\|_{2,S}\vee\delta_n} \le 2u\Big) \ge 1-\frac{1}{4u^2} \ge 3/4,$$

for all $u\ge 1$. Use a symmetrization lemma of Pollard (1984, Lemma II.3.8), see Appendix, to obtain

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\frac{\|h\|_{2,S_n}}{\|h\|_{2,S}\vee\delta_n}\ge 14u\Big) \le 2\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\frac{|\|h\|_{2,S_n'}-\|h\|_{2,S_n''}|}{\|h\|_{2,S}\vee\delta_n}\ge 12u\Big) .$$

The peeling device on the set

$$\{h\in\mathcal{H} : (2u)^{j-1}\delta_n \le \|h\|_{2,S} \le (2u)^j\delta_n\ , j=1,2,\dots\}$$

and the inequality in Pollard (1984, page 33) give

$$\mathbb{P}\Big(\sup_{h\in\mathcal{H}}\frac{|\|h\|_{2,S_n'}-\|h\|_{2,S_n''}|}{\|h\|_{2,S}\vee\delta_n}\ge 12u \;\Big|\; Z_1,\dots,Z_n\Big)$$

$$\le \sum_{j=1}^\infty \mathbb{P}\Big(\sup_{\substack{h\in\mathcal{H}\\ \|h\|_{2,S}\le(2u)^j\delta_n}} |\|h\|_{S_n'}-\|h\|_{S_n''}| \ge 6(2u)^j\delta_n \mid Z_1,\dots,Z_n\Big)$$

$$\le \sum_{j=1}^\infty 2\exp\Big(H(\sqrt{2}(2u)^j\delta_n,\mathcal{H},\bar{S}_{2n}) - 2n(2u)^{2j}\delta_n^2\Big)$$

$$\le \sum_{j=1}^\infty 2\exp\Big(H((2u)^j\delta_n,\mathcal{H},S_n{}') + H((2u)^j\delta_n,\mathcal{H},S_n{}'') - 2n(2u)^{2j}\delta_n^2\Big)$$

$$\leq \sum_{j=1}^{\infty} 2\exp\Big(-n(2u)^{2j}\delta_n^2\Big)\,, \tag{14}$$

where the last inequality is obtained using that since $n\delta_n^2 \geq 2A_oH(\delta_n)$, also $nt^2 \geq 2A_oH(t)$ for all $t \geq \delta_n$ (here $t = (2u)^j\delta_n$). Observe that, since $(2u)^{2j} \geq (2u)^2 j > u^2 j$ for all $u \geq 1$ and $j \geq 1$, we have

$$\sum_{j=1}^{\infty} \exp(-n(2u)^{2j}\delta_n^2) \leq 2\exp(-u^2 n\delta_n^2)\,, \tag{15}$$

whenever $n\delta_n^2 > \log 2$. We finish the proof by combining (14) and (15), and taking $\delta_n = n^{-\frac{1}{2+2\rho}}$. ■

Appendix A.

Proof of Lemma 1. We write $L(f(x)) = \mathbb{E}_{Y|X}[l(Y, f(X))|X = x]$ and recall that $p_j(x) = P(Y = j|X = x)$ for all $j = 1,\ldots,m$, and that $f = (f_1,\ldots,f_m)$ with $\sum_{j=1}^m f_j = 0$. Definition (4) of the loss and the fact that $\sum_{j=1}^m p_j = 1$ give

$$L(f) = \sum_{j=1}^m p_j\Big(\sum_{k=1,\,k\neq j}^m (f_k + \frac{1}{m-1})_+\Big) = \sum_{j=1}^m (1-p_j)(f_j + \frac{1}{m-1})_+\,.$$

Let $p_k = \max_{j\in\{1,\ldots,m\}} p_j$. Here $f_j^* = -1/(m-1)$ for all $j \neq k$, and $f_k^* = 1$. Let $J^+(k) = \{j \neq k\,:\, f_j \geq -1/(m-1)\,,\ j = 1,\ldots,m\}$ and $J^-(k) = \{j \neq k\,:\, f_j < -1/(m-1)\,,\ j = 1,\ldots,m\}$. Write

$$\begin{aligned}\Delta(f) &:= L(f) - L(f^*)\\ &= \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+ \,+\, (1-p_k)(f_k + \frac{1}{m-1})_+ \,-\, (1-p_k)(1+\frac{1}{m-1})\,.\end{aligned}$$

We first consider the case $f_k \geq -1/(m-1)$. Here,

$$\Delta(f) = (1-p_k)(f_k - 1) \,+\, \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+\,.$$

The zero-sum constraint $\sum_{j=1}^m f_j = 0$ simply implies $f_k - 1 = -\sum_{j\neq k}(f_j + \frac{1}{m-1})$. Divide the sum into the sets $J^+(k)$ and $J^-(k)$ to obtain

$$\Delta(f) = \sum_{j\in J^+(k)} (p_k - p_j)\,(f_j + \frac{1}{m-1}) \,+\, (1-p_k)\sum_{j\in J^-(k)} |f_j + \frac{1}{m-1}|\,.$$

For the case $f_k < -1/(m-1)$, observe that

$$\frac{m}{m-1} \;=\; \sum_{j\neq k}(f_j + \frac{1}{m-1}) + f_k + \frac{1}{m-1} \;<\; \sum_{j\neq k}(f_j + \frac{1}{m-1})$$

to obtain

$$\begin{aligned}\Delta(f) &= (1-p_k)\,(-\frac{m}{m-1}) + \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+\\ &> (p_k - 1)\sum_{j\neq k}(f_j + \frac{1}{m-1}) + \sum_{j\neq k}(1-p_j)(f_j + \frac{1}{m-1})_+\\ &= \sum_{j\in J^+(k)} (p_k - p_j)\,(f_j + \frac{1}{m-1}) \,+\, (1-p_k)\sum_{j\in J^-(k)} |f_j + \frac{1}{m-1}|\,.\end{aligned}$$

In both cases clearly $L(f)-L(f^*)$ is always non-negative since $p_k - p_j$ is non-negative for all $j \neq k$. It follows that

$$R(f)-R(f^*) = \sum_{k=1}^{m} \int (L(f)-L(f^*))\ \mathbb{1}(p_k = \max_{j=1,\ldots,m} p_j)\ dQ$$

is always non-negative, with Q the unknown marginal distribution of X. ■

Proof of Lemma 3. Let τ be defined as in (8). We write $L(f(x)) = \mathbb{E}_{Y|X}[l(Y, f(X))|X = x]$ and recall that $p_j(x) = P(Y = j|X = x)$ for all $j = 1,\ldots,m$, and that $f = (f_1,\ldots,f_m)$ with $\sum_{j=1}^m f_j = 0$. From the proof of Lemma 1, clearly

$$(L(f)-L(f^*))\ \mathbb{1}(p_k = \max_{j=1,\ldots,m} p_j) \geq \tau \sum_{j\neq k} |f_j - f_j^*| \geq \frac{\tau}{2} \sum_{j=1}^{m} |f_j - f_j^*|\ ,$$

where the second inequality is obtained from the fact that $|f_k - f_k^*| \leq \sum_{j\neq k} |f_j - f_j^*|$. That is, the excess risk is lower bounded by

$$\frac{1}{2} \sum_{j=1}^{m} \int \tau |f_j - f_j^*| dQ\ .$$

It implies that, for all $z > 0$,

$$R(f) - R^* \geq \frac{z}{2} \sum_{j=1}^{m} \left[\int |f_j - f_j^*| dQ - \int_{\tau \leq z} |f_j - f_j^*| dQ \right]\ .$$

Since $|f_j - f_j^*| \leq M$ for all j, and by Condition AA, the second integral in the inequality above can be upper bounded by $M(Cz)^{1/\gamma}$. Thus, for all $z > 0$,

$$R(f) - R^* \geq \frac{z}{2} \sum_{j=1}^{m} \int |f_j - f_j^*| dQ\ -\ \frac{z}{2} m M (Cz)^{1/\gamma}\ .$$

We take $z = \left(\sum_{j=1}^m \int |f_j - f_j^*| dQ \right)^\gamma / \left(mMC^{1/\gamma}(1+\gamma^{-1}) \right)^\gamma$ when $\gamma > 0$, and $z \uparrow 1/C$ when $\gamma = 0$. ■

Symmetrization lemma (Pollard, 1984, Lemma II.3.8)**.** *Let $\{Z(t) : t \in T\}$ and $\{Z'(t) : t \in T\}$ be independent stochastic process sharing an index set T. Suppose there exist constants $\beta > 0$ and $\alpha > 0$ such that $\mathbb{P}(|Z(t)| \leq \alpha) \geq \beta$ for every $t \in T$. Then*

$$\mathbb{P}\left(\sup_t |Z(t)| > \varepsilon \right) \leq \beta^{-1} \mathbb{P}\left(\sup_t\ |Z(t) - Z'(t)| > \varepsilon - \alpha \right)\ .$$

References

Peter L. Bartlett and Marten H. Wegkamp. Classification with a reject option using a hinge loss. Technical report, U.C. Berkeley, 2006.

Peter L. Bartlett, Michael I. Jordan, and Jon D. McAuliffe. Convexity, classification and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM: Probability and Statistics*, 9:323–375, 2005.

Koby Crammer and Yoram Singer. On the learnability and design of output codes for multiclass problems. In *Proceeding of the 13th Annual Conference on Computational Learning Theory*, pages 35–46. Morgan Kaufmann, 2000.

Koby Crammer and Yoram Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *Journal of Machine Learning Research*, 2:265–292, 2001.

Eustasio del Barrio, Paul Deheuvels, and Sara A. van de Geer. *Lectures on Empirical Processes*. EMS Series of Lectures in Mathematics. European Mathematical Society, 2007.

Kaibo Duan and S. Sathiya Keerthi. Which is the best multiclass svm method? an empirical study. In *Multiple Classifier Systems*, number 3541 in Lecture Notes in Computer Science, pages 278–285. Springer Berlin/Heidelberg, 2005.

Theodoros Evgeniou, Massimiliano Pontil, and Tomaso Poggio. Regularization networks and support vector machines. *Advances in Computational Mathematics*, 13:1–50, 2000.

Yann Guermeur. Combining discriminant models with new multiclass svms. *Pattern Analysis & Applications*, 5:168–179, 2002.

Godfrey H. Hardy, John E. Littlewood, and George Pólya. *Inequalities*. Cambridge University Press, second edition, 1988.

Chih-Wei Hsu and Chih-Jen Lin. A comparison methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425, 2002.

Yoonkyung Lee. *Multicategory Support Vector Machines, Theory and Application to the Classification of Microarray Data and Satellite Radiance Data*. PhD thesis, University of Wisconsin-Madison, Departement of Statistics, 2002.

Yoonkyung Lee and Zhenhuan Cui. Characterizing the solution path of multicategory support vector machines. *Statistica Sinica*, 16(2):391–409, 2006.

Yoonkyung Lee, Yi Lin, and Grace Wahba. Multicategory support vector machines: Theory and application to the classification of microarray data and satellite radiance data. *Journal of the American Statistical Association*, 99(465):67–81, 2004.

Yi Lin. Support vector machines and the bayes rule in classification. *Data Mining and Knowledge Discovery*, 6(3):259–275, 2002.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6): 1808–1829, 1999.

David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag New York Inc., 1984.

Shuguang Song and Jon A. Wellner. An upper bound for uniform entropy numbers. Technical report, Departement of Statistics, University of Washington, 2002. URL `www.stat.washington.edu/www/research/reports/#2002/tr409.ps`.

Ingo Steinwart and Clint Scovel. Fast rates for support vector machines. In P. Auer and R. Meir, editors, *COLT*, volume 3559 of *Lecture Notes in Computer Science*, pages 279–294, 2005.

Bernadetta Tarigan and Sara A. van de Geer. Classifiers of support machine type with l_1 complexity regularization. *Bernoulli*, 12(6):1045–1076, 2006.

Ambuj Tewari and Peter L. Bartlett. On the consistency of multiclass classification methods. In P. Auer and R. Meir, editors, *COLT*, volume 3559 of *Lecture Notes in Computer Science*, pages 143–157, 2005.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Annals of Statistics*, 32:135–166, 2004.

Sara A. van de Geer. *Empirical Processes in M-estimation*. Cambridge University Press, 2000.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996.

Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 2000.

Lifeng Wang and Xiaotong Shen. On l_1-norm multiclass support vector machines: Methodology and theory. *Journal of the American Statistical Association*, 102(478):583–594, 2007.

Jason Weston and Chris Watkins. Multi-class support vector machines. In *Proceedings of ESANN99*, 1999.

Tong Zhang. Statistical behaviour and consistency of classification methods based on convex risk minimization. *The Annals of Statistics*, 32(1):56–134, 2004a. With discussion.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004b.

Tong Zhang. Statistical analysis of some multi-category large margin classification methods. *Journal of Machine Learning Research*, 5:1225–1251, 2004c.

Hui Zou, Ji Zhu, and Trevor Hastie. The margin vector, admissible loss and multi-class margin-based classifiers. Technical report, Statistics Departement, Stanford University, 2006.