Oracle inequalities for multi-fold cross validation

Vaart, A.W. van der; Dudoit, S.; Laan, M.J.

Statistics & Decisions , 1– (2006)

Oracle Inequalities for Multi-fold Cross Validation

A.W. van der Vaart, S. Dudoit, M.J. van der Laan

Received: Month-1 99, 2003; Accepted: Month-2 99, 2004

Summary: We consider choosing an estimator or model from a given class by cross validation consisting of holding a nonneglible fraction of the observations out as a test set. We derive bounds that show that the risk of the resulting procedure is (up to a constant) smaller than the risk of an oracle plus an error which typically grows logarithmically with the number of estimators in the class. We extend the results to penalized cross validation in order to control unbounded loss functions. Applications include regression with squared and absolute deviation loss and classification under Tsybakov's condition.

1 Introduction

Let $X_1, \ldots, X_n$ be a sample of observations, independent and identically distributed random variables, distributed according to a probability measure P on a measurable space $(\mathcal{X}, \mathcal{A})$. For a given parameter set Θ and "loss function" $L\colon \mathcal{X} \times \Theta \to [0, \infty)$ we aim at finding an estimator $\hat{\theta}$ that minimizes the function $R\colon \Theta \to \mathbb{R}$ defined by

$$R(\theta) = \int L(x, \theta)\, dP(x) = \mathrm{E} L(X_1, \theta). \tag{1.1}$$

Here an "estimator" $\hat{\theta}$ is as usual a measurable function of the observations and $x \mapsto L(x, \theta)$ is assumed measurable. A proper statistical setting would require to consider the "(prediction) risk" R also as a function of the unknown distribution P, but we do not make this explicit in the notation as in the results of this paper only a single "true" distribution P appears.

For notational convenience we assume that the estimator is defined for each n and symmetric in the observations, so that it can be written as a function $\hat{\theta} = \theta(\mathbb{P})$, for $\mathbb{P} = n^{-1}\sum_{i=1}^n \delta_{X_i}$ the empirical distribution of the observations and θ a map from the set of uniform discrete distributions into the parameter set. Given a collection $\{\theta_k(\mathbb{P})\colon k \in \mathcal{K}\}$ of estimators we wish to select the estimator $\theta_{\hat{k}}(\mathbb{P})$ that minimizes R, where $\hat{k}$ may itself depend on the observations. Because R depends on the unknown distribution P,

AMS 1991 subject classification: Primary: 62G15, 62G20, 62F25
Key words and phrases: Model selection, oracle inequality, adaptation

this cannot be achieved exactly. However, we try and approximate our aim by cross validation, as follows.

We split the the data randomly into two sets, a *training* and a *test* (or *validation*) sample. To formalize this let $S = (S_1, \ldots, S_n)$ be a random vector independent of $X_1, \ldots, X_n$ and taking values in $\{0,1\}^n$. If $S_i = 0$, then X_i belongs to the first (training) subset; otherwise it belongs to the second (test) subset. Define sub-empirical distributions $\mathbb{P}_S^0$ and $\mathbb{P}_S^1$ by

$$\mathbb{P}_S^j = \frac{1}{n^j} \sum_{i: S_i = j} \delta_{X_i}, \qquad n^j = \#(1 \le i \le n: S_i = j), \qquad j = 0, 1.$$

The "randomness" of the split is actually of no importance in the following: the split may be deterministic. The only assumption is that S is stochastically independent of the observations. Given a collection $\{\theta_k: k \in \mathcal{K}\}$ of estimators we form candidate estimates $\theta_k(\mathbb{P}_S^0)$ by applying the estimators to the training sample. The risk of these estimators, averaged over the splits, as a function of $k \in \mathcal{K}$, is equal to

$$k \mapsto \mathrm{E}_S \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, dP(x) = \mathrm{E}_S R\big(\theta_k(\mathbb{P}_S^0)\big). \tag{1.2}$$

The value $\tilde{k} \in \mathcal{K}$ that minimizes this expression depends on the observations as well as on the unknown distribution P, and hence is unavailable. In view of the latter it is referred to as an *oracle*. Cross validation replaces P by $\mathbb{P}_S^1$ and proposes to use the value $\hat{k}$ that minimizes

$$k \mapsto \mathrm{E}_S \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\mathbb{P}_S^1(x). \tag{1.3}$$

Next the final estimator is $\mathrm{E}_S \theta_{\hat{k}}(\mathbb{P}_S^0)$ or perhaps $\theta_{\hat{k}}(\mathbb{P})$.

Example 1.1 (regression). In the regression model the observations are a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in a space $\mathcal{X} \times \mathbb{R}$ and satisfying a model

$$Y = \theta_0(X) + \varepsilon,$$

for ε an unobservable "error". The purpose is to estimate the function $\theta_0: \mathcal{X} \to \mathbb{R}$.

To fit this in the preceding set-up we take the pairs (X_i, Y_i) as the observations (rather than the X_i), and the parameter set Θ as a collection of functions $\theta: \mathcal{X} \to \mathbb{R}$. A popular loss function in this setting is the squared error loss

$$L\big((x,y),\theta\big) = \big(y - \theta(x)\big)^2.$$

If the conditional mean of the error given X is zero, then the corresponding risk is $R(\theta) = \mathrm{E}\varepsilon^2 + \mathrm{E}(\theta - \theta_0)^2(X)$, so that minimizing R is equivalent to estimation of θ_0 under L_2-loss.

An alternative is to replace the square by another increasing function of the discrepancy $\big|y - \theta(x)\big|$. It may not be possible to express the risk then in a simple distance on the regression functions, but it can always be understood in terms of prediction error.

Example 1.2 (classification). In the classification model the observations are a sequence of pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ taking values in a space $\mathcal{X} \times \{0, 1\}$. The purpose is to predict the value of a future outcome $Y \in \{0, 1\}$ from a given future input X, where (X, Y) is distributed as the observations. A "classifier" is a measurable function $\theta\colon \mathcal{X} \to \{0, 1\}$, and a natural loss function is

$$L\big((x, y), \theta\big) = 1_{y \neq \theta(x)}.$$

The corresponding risk is the probability $R(\theta) = \mathrm{P}\big(Y \neq \theta(X)\big)$ that the classifier fails to predict the outcome correctly. Relative to all possible classifiers this is minimized by the *Bayes classifier* $\theta_0 = 1_{\eta_0 \geq 1/2}$ for $\eta_0(x) = \mathrm{P}(Y = 1 | \, X = x)$, and we may view the problem also as aimed at estimating θ_0 under this loss.

Example 1.3 (multivariate mean). Suppose that we observe a sample $X_1, \ldots, X_n$ from a D-variate normal distribution with mean θ_0 and covariance matrix the identity matrix, and we wish to estimate the mean vector $\theta_0 \in \mathbb{R}^D$ relative to the loss function

$$L(x, \theta) = \|x - \theta\|^2.$$

The corresponding risk function $R(\theta) = \|\theta - \theta_0\|^2 + D$ is essentially the square Euclidean distance.

By sufficiency this problem is equivalent to estimating D univariate means $\theta_1, \ldots, \theta_D$ each based on a single $N(\theta_i, 1/n)$-observation. The vector of sample means is an obvious estimator, but may be unattractive if D is large, when shrinkage estimators perform a better job, and a-priori information, for instance sparsity of the vector θ, may suggest many other estimators. Cross validation can be used to choose from these estimators.

For fixed values of the training sample $\mathbb{P}_S^0$ the expression (1.3) is an unbiased estimate of its mean, which is the risk $R\big(\theta_k(\mathbb{P}_0^S)\big) = \int L\big(x, \hat{\theta}_k(\mathbb{P}_S^0)\big)\, dP(x)$ of $\theta_k(\mathbb{P}_S^0)$. We may expect that the k which minimizes the estimated risk (1.3) will also approximately minimize this population risk. The realization of this expectation depends on the quality (i.e. variability) of the risk estimate, next to its being unbiased. In the next sections we present inequalities that show that $\hat{k}$ indeed (nearly) minimizes the risk. More precisely we show that the risk of the estimator given by $\hat{k}$ is not much bigger than the risk of the "oracle estimator", which uses $\tilde{k}$ defined as the minimizer of (1.2). This is achieved by comparing the deviation of (1.3) from its mean using inequalities from empirical process theory.

The quality of the risk estimates is determined by the number and type of procedures θ_k. It is also dependent on the number of observations in the test sample. In practice two-fold, three-fold and ten-fold splits appear popular. The leave-one-out cross-validation scheme ("n-fold cross validation", see Stone (1974, 1976)) averages over all validation sets consisting of one observation. This scheme also leads to unbiased risk estimates, but the methods of this paper cannot be used to analyze the resulting procedure. Leave-one-out cross validation is studied theoretically in Li (1987) and Andrews (1991) for nonparametric regression, and as recently as in Davies et al. (2005) for Kullback-Leibler divergence in parametric models.

Choosing a best estimator from a given set is also known as "model selection" and has recently been studied within the context of aggregation of estimators. For instance, linear aggregation proposes the best linear combination of the estimators, where the weights may be dependent on the observations. Aggregation has been studied, among others, in Nemirovski (2000), Yang (2000), Bunea et al. (2004), and Tsybakov (2004), and appears to be a powerful technique. Also see George (2000) and the references cited there for further connections to model selection.

Through its focus on risk estimation, cross validation is connected to penalized contrast estimation. For given numbers $\lambda(k,\theta)$ the latter procedure selects the estimator $\theta_{\hat{k}}(\mathbb{P})$ for $\hat{k}$ the minimizer of

$$k \mapsto \int L\big(x, \theta_k(\mathbb{P})\big)\, d\mathbb{P}(x) + \frac{\lambda\big(k, \theta_k(\mathbb{P})\big)}{n}. \tag{1.4}$$

The *penalty* $\lambda(k,\theta)/n$ is meant to prevent overfitting the data by making complex estimators less favorable. Alternatively, a penalty can be understood (at least partly) as a correction for the double use of the data in (1.4). The empirical integral in the display is meant to estimate the population integral $\int L\big(x, \theta_k(\mathbb{P})\big)\, dP(x)$. However, the mean of the empirical integral is $\mathrm{E}L\big(X_1, \theta_k(\mathbb{P})\big)$, in which the variable X_1 appears twice, once as the first argument of L and a second time hidden in $\mathbb{P}$. Typically $\mathrm{E}L\big(X_1, \theta_k(\mathbb{P})\big)$ is smaller than $\mathrm{E}\int L\big(x, \theta_k(\mathbb{P})\big)\, dP(x)$ and hence the empirical integral in (1.4) underestimates $\int L\big(x, \theta_k(\mathbb{P})\big)\, dP(x)$. The penalty (called "covariance correction" by Efron (2004)) is added to remedy this. The link between penalties and risk estimation was made by Akaike (1973, 1974) and Mallows (1973). Oracle inequalities were obtained by Li (1987), Barron and Cover (1991), Vapnik (1998), Barron et al. (1999), Lugosi and Nobel (1999), Massart (2000), Koltchinskii (2001), van de Geer (2001), Wegkamp (2003), and Birgé (2006), among many others. See Boucheron et al. (2005) for a review in the context of the classification problem.

The advantage of penalized contrast estimation is that it is computationally more efficient, and avoids sampling splitting, thus referring directly to the estimator based on all observations. The disadvantage is that appropriate penalties must be worked out for each situation at hand. The latter may be complicated and may lead to suboptimal estimators.

The cross validation procedure uses independent observations to construct the estimators $\theta_k(\mathbb{P}_S^0)$ and to estimate the risk $\int L\big(x, \theta_k(\mathbb{P}_0^S)\big)\, dP(x)$ (using $\mathbb{P}_S^1$). Thus it provides an unbiased estimate of risk, and a penalty seems unnecessary. Nevertheless in Section 3 we consider the combination of cross validation and penalization. The introduction of penalties does not complicate the situation much, and penalization appears to be potentially useful to control the variance of the risk estimator. In particular, for unbounded loss functions the risk estimator (1.3), even though unbiased, may become imprecise due to a large variance. A penalty can help to downweight estimators whose risk is difficult to estimate. This is illustrated for regression and the multivariate mean problem in Sections 4 and 7.

Although k-fold cross validation is applied routinely, oracle inequalities of the type of this paper appear to have been first obtained in Devroye and Lugosi (2001), Györfi

et al. (2002), and Dudoit and van der Laan (2005). (In an earlier paper Zhang (1993) studied the distribution of a selector among finitely many models.) The contribution of the present paper is to refine and extend the results in these papers. The main result allows unbounded errors (e.g. Gaussian) and loss functions in the regression model and covers the classification model under Tsybakov's condition. Furthermore, we introduce penalties to cover unbounded regression functions.

We use the notation Pf for the integral $\int f\,dP$ of a function relative to a measure P. Furthermore $\mathbb{P}$ is the empirical measure and $\mathbb{G} = \sqrt{n}(\mathbb{P} - P)$ is the empirical process of the n observations $X_1, \ldots, X_n$, and we write $\mathbb{G}f = \sqrt{n}(\mathbb{P}f - Pf)$. Similarly, the empirical processes corresponding to the subsamples are $\mathbb{G}_S^j = \sqrt{n^j}(\mathbb{P}_S^j - P)$. For notational convenience we let X be a random variable independent of $X_1, \ldots, X_n$ with the same distribution. We assume throughout that the size n^1 of the test sample is bounded below by a positive constant times n. We write $a \lesssim b$ if $a \le Cb$ for a constant C that is fixed within the context.

Theorem 2.3 in Section 2 is the main oracle inequality, which is extended to include penalties in Theorem 3.2 in Section 3. Sections 4, 5, 6 and 7 apply these results to regression and classification, with an application to adaptive estimation in Section 4. Section 8 contains most of the proofs.

2 Oracle inequalities

Let $\hat{k}$ and $\tilde{k}$ be the minimizers of (1.3) and (1.2), respectively. The purpose is to show that $\hat{k}$ yields a risk that is not much bigger than the risk provided by the oracle $\tilde{k}$.

From the minimizing property of $\hat{k}$ it is immediate that

$$\mathrm{E}_S \int L\big(x, \theta_{\hat{k}}(\mathbb{P}_S^0)\big)\, d\mathbb{P}_S^1(x) \le \mathrm{E}_S \int L\big(x, \theta_{\tilde{k}}(\mathbb{P}_S^0)\big)\, d\mathbb{P}_S^1(x). \tag{2.1}$$

If we replace the empirical measure $\mathbb{P}_S^1$ by the true distribution P, then we make an error that can be expressed in the empirical process $\mathbb{G}_S^1$. This leads to the following basic lemma.

Lemma 2.1 *For any $\delta > 0$,*

$$\begin{aligned}\mathrm{E}_S \int L\big(x, \theta_{\hat{k}}(\mathbb{P}_S^0)\big)\, dP(x) &\le (1+2\delta)\,\mathrm{E}_S \int L\big(x, \theta_{\tilde{k}}(\mathbb{P}_S^0)\big)\, dP(x) \\ &\quad + \frac{1}{\sqrt{n^1}} \mathrm{E}_S \max_{k\in\mathcal{K}} \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\big((1+\delta)\mathbb{G}_S^1 - \delta\sqrt{n^1}P\big)(x) \\ &\quad + \frac{1}{\sqrt{n^1}} \mathrm{E}_S \max_{k\in\mathcal{K}} \int -L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\big((1+\delta)\mathbb{G}_S^1 + \delta\sqrt{n^1}P\big)(x).\end{aligned}$$

Proof: By simple algebra the minimizing property (2.1) can be written in the form

$$\begin{aligned}\mathrm{E}_S \int L\big(x, \theta_{\hat{k}}(\mathbb{P}_S^0)\big)\, dP(x) &\le (1+2\delta)\mathrm{E}_S \int L\big(x, \theta_{\tilde{k}}(\mathbb{P}_S^0)\big)\, dP(x)\\ &+\frac{1}{\sqrt{n^1}}\mathrm{E}_S \int L\big(x, \theta_{\tilde{k}}(\mathbb{P}_S^0)\big)\, d\big((1+\delta)\mathbb{G}_S^1 - \delta\sqrt{n^1}P\big)(x)\\ &-\frac{1}{\sqrt{n^1}}\mathrm{E}_S \int L\big(x, \theta_{\hat{k}}(\mathbb{P}_S^0)\big)\, d\big((1+\delta)\mathbb{G}_S^1 + \delta\sqrt{n^1}P\big)(x).\end{aligned}$$

We can next replace the two random variables $\hat{k}$ and $\tilde{k}$ by the maximum over $k \in \mathcal{K}$. □

The idea is that the second and third terms on the right in the lemma are very small, as they are preceded by $(n^1)^{-1/2}$ and concern the empirical process $\mathbb{G}_S^1$, which is centered (and shifted downward or upward if $\delta > 0$). If the two terms are negligible, then the lemma asserts that the cross-validated estimator, given by $\hat{k}$, has a risk that is at most $1 + 2\delta$ times the risk of the oracle estimator given by $\tilde{k}$. The choice $\delta = 0$ gives the best comparison of cross validation and oracle risk, but it will be seen that this choice comes at the price that the two remainder terms are larger. This is because for $\delta > 0$ these terms involve the decentered empirical processes $(1+\delta)\mathbb{G}_S^1 - \delta\sqrt{n^1}P$ and $(1+\delta)\mathbb{G}_S^1 + \delta\sqrt{n^1}P$ in the remainder term. Pulling the variables in the maximum away from their expectation can have a dramatic effect on their expected value.

It is relatively easy to make this idea precise. Given the split S and the observations $\mathbb{P}_S^0$ in the first set of observations, the empirical process $\mathbb{G}_S^1$ is an ordinary stochastic process based on $n^1 = \#(S_i = 1)$ observations. We can therefore apply any maximal inequality for empirical processes to find a bound on the expectation of the right side of the lemma given S and $\mathbb{P}_S^0$. For instance, in case that $\delta = 0$, we write, with E_Z meaning "expectation relative to the variable Z",

$$\mathrm{E} \max_{k\in\mathcal{K}} \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\mathbb{G}_S^1(x) = \mathrm{E}_{S,\mathbb{P}_S^0}\mathrm{E}_{\mathbb{P}_S^1} \max_{k\in\mathcal{K}} \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\mathbb{G}_S^1(x),$$

and apply maximal inequalities to the inner expectation on the right, for fixed S and $\mathbb{P}_S^0$. This will typically show that the "remainder terms" on the right in the preceding lemma are of the order $n^{-1/2}$ times an expression involving the complexity of the set of estimators. If the distributions of the losses $L(X, \theta)$ have exponential tails, then the cardinality $\#\mathcal{K}$ of the set of estimators will typically enter at most logarithmically, giving oracle inequalities of the type, for some $p > 0$,

$$\mathrm{E} \int L\big(x, \theta_{\hat{k}}(\mathbb{P}_S^0)\big)\, dP(x) \le \mathrm{E} \int L\big(x, \theta_{\tilde{k}}(\mathbb{P}_S^0)\big)\, dP(x) + O\Big(\frac{(\log \#\mathcal{K})^{1/p}}{\sqrt{n}}\Big). \quad (2.2)$$

We conclude that the empirical choice $\hat{k}$ results in a risk that is at most a constant times $(\log \#\mathcal{K})^{1/p}/\sqrt{n}$ bigger than the risk obtained by the oracle $\tilde{k}$.

If the loss functions are uniformly bounded, then it is particularly easy to make this precise. For instance, writing the empirical process of an i.i.d. sample of size n as $\mathbb{G}$, we have for any set $\mathcal{F}$ of (bounded) measurable functions (assume $\#\mathcal{F} \ge 2$)

$$\mathrm{E} \max_{f\in\mathcal{F}} |\mathbb{G}f| \lesssim \frac{\log \#\mathcal{F}}{\sqrt{n}} \max_{f\in\mathcal{F}} \|f\|_\infty + \sqrt{\log \#\mathcal{F}} \max_{f\in\mathcal{F}} \|f\|_2, \quad (2.3)$$

(e.g. van der Vaart and Wellner (1996), formula (2.5.5)). Similar bounds are valid for certain unbounded functions. For instance, assume that the functions f possess exponentially decreasing tails of order p: for some constant $M(f)$ and every $t > 0$,

$$P\big(x\colon |f(x)| > t\big) \lesssim e^{-t^p/M(f)^p}.$$

Then for $1 \leq p \leq 2$ the variables $\mathbb{G}f$ possess tails of the same order (see page 245 in van der Vaart and Wellner (1996)), and hence (e.g. van der Vaart and Wellner (1996), Lemmas 2.2.2 and 2.2.1)

$$\mathrm{E} \max_{f \in \mathcal{F}} |\mathbb{G}f| \lesssim (\log \#\mathcal{F})^{1/p} \max_{f \in \mathcal{F}} M(f). \tag{2.4}$$

In particular, if the functions f are bounded, then we can take $p = 2$ and $M(f)$ equal to a multiple of $\|f\|_\infty$, in view of Hoeffding's inequality. Alternatively, if $\big(M(f), v(f)\big)$ are Bernstein pairs for the functions f (see the definition below), then (2.3) holds but with $\|f\|_\infty$ replaced by $M(f)$ and $\|f\|_2$ replaced by $v(f)$ (van der Vaart and Wellner (1996), Lemma 3.4.3; or see the appendix for more general results).

Bounds of the type (2.2) are of interest only if the remainder $O\big((\log \#\mathcal{K})^{1/p}/\sqrt{n}\big)$ is of smaller order than the oracle risk. This is not always the case. For instance, in the regression situation with square error loss, the oracle risk may well be of order $O(1/n)$ if one of the estimators corresponds to a finite-dimensional model that contains the true regression function, and it will also be much smaller than $n^{-1/2}$ in the situation of not too large nonparametric models. Such fast rates are also possible in classification problems where the Bayes classifier does not concentrate too much near $1/2$ (see Mammen and Tsybakov (1999)). We can obtain alternative bounds where the $n^{-1/2}$ is replaced by n^{-1} at the price of choosing δ positive.

Given a measurable function $f\colon \mathcal{X} \to \mathbb{R}$, call $\big(M(f), v(f)\big)$ a pair of *Bernstein numbers* of f if

$$M(f)^2\, P\Big(e^{|f|/M(f)} - 1 - \frac{|f|}{M(f)}\Big) \leq \tfrac{1}{2} v(f).$$

It may be shown (see Section 8.1) that:

(i) If f is uniformly bounded, then $\big(\|f\|_\infty, 1.5Pf^2\big)$ is a pair of Bernstein numbers.

(ii) If $|f| \leq g$, then a Bernstein pair for g is also a Bernstein pair for f.

(iii) If $\big(M(f), v(f)\big)$ and $\big(M(g), v(g)\big)$ are Bernstein pairs for f and g, then $2\big(M(f) \vee M(g), v(f) + v(g)\big)$ is a Bernstein pair for $f + g$.

(iv) If $\big(M(f), v(f)\big)$ is a Bernstein pair for f and $c > 0$, then $\big(cM(f), c^2 v(f)\big)$ is a Bernstein pair for cf.

In view of (i) the numbers $M(f)$ and $v(f)$ could be intuitively thought of as "supremum" and "variance" of f. However, the usefulness of Bernstein numbers goes beyond this example. The other properties roughly show that Bernstein pairs behave as supremum and variance under simple operations. Because we use Bernstein pairs to control the variables

$\mathbb{G}f$ and $\mathbb{G}(f+c) = \mathbb{G}f$ for every constant c, throughout Bernstein pairs $\big(M(f), v(f)\big)$ can be replaced by pairs $\big(M(f+c), v(f+c)\big)$ for every constant c.

The following maximal inequality is a consequence of Lemma 8.2 in Section 8 (with $q = 1$).

Lemma 2.2 *Let $\mathbb{G}$ be the empirical process of an i.i.d. sample of size n from the distribution P and assume that $Pf \geq 0$ for every $f \in \mathcal{F}$. Then, for any Bernstein pairs $\big(M(f), v(f)\big)$ and for any $\delta > 0$ and $1 \leq p \leq 2$,*

$$\mathrm{E} \max_{f \in \mathcal{F}} (\mathbb{G} - \delta\sqrt{n}P)f \leq \frac{8}{n^{1/p-1/2}} \log(1 + \#\mathcal{F}) \max_{f \in \mathcal{F}} \Big[\frac{M(f)}{n^{1-1/p}} + \Big(\frac{v(f)}{(\delta Pf)^{2-p}} \Big)^{1/p} \Big].$$

The same upper bound is valid for $\mathrm{E} \max_{f \in \mathcal{F}} (\mathbb{G} + \delta\sqrt{n}P)(-f)$.

Application of Lemma 2.2 to the second and third terms on the right in Lemma 2.1 with the collection $\mathcal{F}$ equal to the functions $x \mapsto L(x, \theta)$ with θ ranging over Θ, yields the following oracle inequality.

Theorem 2.3 *For $\theta \in \Theta$ let $\big(M(\theta), v(\theta)\big)$ be a Bernstein pair for the function $x \mapsto L(x, \theta)$ and assume that $R(\theta) = \int L(x, \theta)\, dP(x) \geq 0$ for every $\theta \in \Theta$. Then for any $\delta > 0$ and $1 \leq p \leq 2$,*

$$\begin{aligned} \mathrm{E} R\big(\theta_{\tilde{k}}(\mathbb{P}_S^0)\big) &\leq (1 + 2\delta)\, \mathrm{E} R\big(\theta_{\tilde{k}}(\mathbb{P}_S^0)\big) + (1 + \delta) \mathrm{E}\Big(\frac{16}{(n^1)^{1/p}} \Big) \\ &\quad \times \log(1 + \#\mathcal{K}) \sup_{\theta \in \Theta} \Big[\frac{M(\theta)}{(n^1)^{1-1/p}} + \Big(\frac{v(\theta)}{R(\theta)^{2-p}} \Big)^{1/p} \Big(\frac{1+\delta}{\delta} \Big)^{2/p-1} \Big]. \end{aligned}$$

In the examples we discuss below the maximum over Θ on the right is finite and hence the remainder term is of the order $O(n^{-1/p})$ times the logarithm of the number of estimators, if the size n^1 of the test sample is a positive fraction of n. For $p = 2$ we can choose $\delta = 0$ and regain the bound of order $O(n^{-1/2})$ obtained in (2.2), albeit that the factor $\log(1 + \#\mathcal{K})$ may not be optimal (cf. (2.4)). For $p = 1$ the bound is of the order $O(n^{-1})$ for every fixed $\delta > 0$.

Because the bound is valid for every $\delta > 0$, in asymptotic applications we can choose $\delta = \delta_n$ tending to zero. Then the oracle inequality can be written in the form $\mathrm{E} R\big(\theta_{\hat{k}}(\mathbb{P}_S^0)\big) \leq \inf_k \mathrm{E} R\big(\theta_k(\mathbb{P}_S^0)\big) + \mathrm{rem}_n$, and an optimal choice of δ_n would make the remainder as small as possible.

The condition that $R(\theta) \geq 0$ can be arranged by defining the loss function L to be centered at its minimum over $\theta \in \Theta$: $L(x, \theta) = L_0(x, \theta) - L_0(x, \theta_0)$ for θ_0 the point of minimum of $\theta \mapsto \int L_0(x, \theta)\, dP(x)$. The cross-validated estimator relative to this centered loss is the same as the cross-validated estimator relative to the original loss (and hence can be implemented without knowledge of θ_0).

The maximum over $\theta \in \Theta$ of the right side of the theorem is bounded only if $v(\theta) \leq D\, R(\theta)^{2-p}$ for every θ and some positive constant D. If $v(\theta)$ is the variance of the function $x \mapsto L(x, \theta)$, then this is true with $p = 1$ if

$$\begin{aligned} R(\theta) = \mathrm{E}\big(L_0(X, \theta) - L_0(X, \theta_0)\big) &\geq d^2(\theta, \theta_0), \\ \mathrm{E}\big(L_0(X, \theta) - L_0(X, \theta_0)\big)^2 &\leq D\, d^2(\theta, \theta_0), \end{aligned} \tag{2.5}$$

for some distance d on Θ and positive constant D. In regular cases the first inequality should be true because θ_0 is a point of minimum, while the second would follow if the loss is Lipschitz in the parameter. The inequality $v(\theta) \leq D\, R(\theta)^{2-p}$ for some $p \in (1, 2]$ corresponds to less regular situations. For instance, in Section 6 it will be seen to be satisfied in the classification problem under Tsybakov's condition.

Lemma 2.2 is based on Bernstein's inequality applied to the variables $\mathbb{G}f$, an exponential tail bound. Alternatively, maximal inequalities for empirical processes may be based on (weaker) moment inequalities on the variables $\mathbb{G}f$, but then the logarithmic factor $\log(1 + \#\mathcal{K})$ will change in a polynomial factor.

The lemma does not exploit relations that may exist between the functions f. If we can control covering numbers (cf. van der Vaart and Wellner (1996)), then we may use more complicated bounds in terms of entropy integrals, which are valid for infinite collections $\mathcal{F}$. In principle an expression such as $\mathrm{E} \max_{f \in \mathcal{F}} |\mathbb{G}f|$ need not grow with the size of $\mathcal{F}$ at all, not even logarithmically. On the other hand, if the estimators θ_k are very different, then not much may be gained from such more involved inequalities.

3 Oracle inequalities with penalties

In this section we combine cross validation with penalization. Given a function $\lambda \colon \mathcal{K} \times \Theta \to [0, \infty)$ the penalized cross-validated estimator is defined as $\theta_{\hat{k}}(\mathbb{P}_S^0)$ for $\hat{k}$ the random element that minimizes, for given observations,

$$k \mapsto \mathrm{E}_S \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, d\mathbb{P}_S^1(x) + \frac{\lambda\big(k, \theta_k(\mathbb{P}_S^0)\big)}{n}. \tag{3.1}$$

The penalized oracle estimator corresponds to the random element $\tilde{k}$ of $\mathcal{K}$ that minimizes

$$k \mapsto \mathrm{E}_S \int L\big(x, \theta_k(\mathbb{P}_S^0)\big)\, dP(x) + \frac{\lambda\big(k, \theta_k(\mathbb{P}_S^0)\big)}{n}.$$

The introduction of penalties is only notationally more involved. We can view it as considering the loss $L(x, \theta) + \lambda(k, \theta)/n$ rather than $L(x, \theta)$, and next apply the results of the preceding section. We restrict ourselves to a particular case: controlling the Bernstein numbers $M(\theta)$ in Theorem 2.3.

The penalties are another source of decentering the variables in the maximum and the minimum, and hence are potentially helpful to control the error term. The decentering takes the form $\delta\sqrt{n}\big(Pf + \lambda(f)/n\big)$ for numbers $\lambda(f)$ rather than $\delta\sqrt{n}Pf$, and can be positive even if $Pf = 0$. The following maximal inequality is a consequence of Lemma 8.2 in Section 8.

Lemma 3.1 *Let $\mathbb{G}$ be the empirical process of an i.i.d. sample of size n from the distribution P, and assume that $Pf + \lambda(f)/n \geq 0$ for every $f \in \mathcal{F}$. Then, for any $\delta > 0$, and*

Bernstein pairs $\big(M(f), v(f)\big)$*, and any* $0 < p \leq 1$ *and* $0 < q \leq 1$*,*

$$\begin{aligned}
&\mathrm{E} \max_{f \in \mathcal{F}} \big(\mathbb{G} - \delta\sqrt{n}P\big)\Big(f + \frac{\lambda(f)}{n}\Big) \\
&\quad \leq \frac{1}{\sqrt{n}}\big[\log(1 + \#\mathcal{F} + D_q)\big]^{1/q} \max_{f \in \mathcal{F}} \Big(\frac{8M(f)}{C_q \delta^{1-q} \lambda(f)^{1-q}}\Big)^{1/q} \\
&\qquad\quad + \frac{1}{\sqrt{n}}\big[\log(1 + \#\mathcal{F} + D_p)\big]^{1/p} \max_{f \in \mathcal{F}} \Big(\frac{8v(f)}{C_p \delta^{2-p} P f \lambda(f)^{1-p}}\Big)^{1/p}.
\end{aligned}$$

Here $C_p > 0$ *and* $D_p \geq 0$ *are constants, equal to* 1 *and* 0 *for* $p = 1$*. The same bound is valid for* $\mathrm{E} \max_{f \in \mathcal{F}} - \big(\mathbb{G} + \delta\sqrt{n}P\big)\big(f + \lambda(f)/n\big)$.

The first maximum on the right is finite if $\lambda(f)^{1-q}$ is proportional to $M(f)$. For the choices $p = q = 1/2$ the right side of the lemma is bounded by a multiple of

$$\frac{1}{\sqrt{n}}\Big[\log(1 + \#\mathcal{F})\Big]^2 \Big[\frac{1}{\delta} \max_{f \in \mathcal{F}} \Big(\frac{M(f)}{\sqrt{\lambda(f)}}\Big)^2 + \frac{1}{\delta^3} \max_{f \in \mathcal{F}} \Big(\frac{v(f)}{Pf\sqrt{\lambda(f)}}\Big)^2\Big].$$

This yields the following theorem.

Theorem 3.2 *For* $\theta \in \Theta$ *let* $\big(M(\theta), v(\theta)\big)$ *be a Bernstein pair for the function* $x \mapsto L(x, \theta)$ *and assume that* $R(\theta) = \int L(x, \theta)\, dP(x) \geq 0$ *for every* $\theta \in \Theta$*. Assume that* $\lambda(k, \theta) = \lambda(\theta)$ *does not depend on* k*. Then, for any* $\delta \in (0, 1)$*, the minimizer* $\hat{k}$ *of (3.1) satisfies, for a universal constant* C*,*

$$\begin{aligned}
\mathrm{E} R\big(\theta_{\hat{k}}(\mathbb{P}_S^0)\big) &\leq (1 + 2\delta)\, \mathrm{E}\Big[R\big(\theta_{\tilde{k}}(\mathbb{P}_S^0)\big) + \frac{\lambda\big(\theta_{\tilde{k}}(\mathbb{P}_S^0)\big)}{n}\Big] \\
&\quad + C\, \mathrm{E} \frac{1}{n^1} \frac{1}{\delta}\big[\log(1 + \#\mathcal{K})\big]^2 \Big[\sup_{\theta \in \Theta} \Big(\frac{M(\theta)}{\sqrt{\lambda(\theta)}}\Big)^2 + \sup_{\theta \in \Theta} \Big(\frac{v(\theta)}{\delta R(\theta)\sqrt{\lambda(\theta)}}\Big)^2\Big].
\end{aligned}$$

A penalty such that $\lambda(\theta) \geq M(\theta)^2$ makes the first maximum on the right finite. Relative to Theorem 2.3 we have then achieved to move the numbers $M(\theta)$ inside the oracle part of the inequality, at the cost of squaring $\log \#\mathcal{K}$. Many variations of this result are possible, also with $\#\mathcal{K} = \infty$ and/or using other penalities (replace Lemma 3.1 by Lemma 8.2). The special choices of the preceding theorem are motivated by the regression model in the next section.

4 Least squares regression

Consider the regression model $Y = \theta_0(X) + \varepsilon$ of Example 1.1, with error with zero conditional mean $\mathrm{E}(\varepsilon \,|\, X) = 0$. The least squares criterion, centered at its minimum, can be written

$$L\big((X, Y), \theta\big) = \big(Y - \theta(X)\big)^2 - \big(Y - \theta_0(X)\big)^2 = 2\varepsilon(\theta_0 - \theta)(X) + (\theta - \theta_0)^2(X).$$

The first term on the right has mean zero, whence the risk is given by

$$R(\theta) = \mathrm{E}L\big((X,Y),\theta\big) = \|\theta - \theta_0\|^2,$$

where $\|\cdot\|$ denotes the L_2-norm relative to the marginal distribution of X. We assume that the error ε has exponential tails, conditionally on X: setting $r_t(X) = \mathrm{E}(e^{t|\varepsilon|}\,|\,X)$, we assume that the function r_t is finite and bounded for some $t > 0$.

Lemma 4.1 *If the regression functions $\theta \in \Theta$ are bounded and the error distribution has conditionally exponential tails, then $\big(M(\theta), v(\theta)\big)$ for*

$$\begin{aligned} M(\theta) &= 4(t^{-1} \vee 1)\big(\|\theta - \theta_0\|_\infty^2 \vee 1\big), \\ v(\theta) &= 2\|\theta - \theta_0\|^2\big(e\|\theta - \theta_0\|_\infty^2 + 8t^{-2}\|r_t\|_\infty\big), \end{aligned}$$

is a Bernstein pair for the function $x \mapsto L(x,\theta)$. This pair satisfies $v(\theta) \lesssim M(\theta)R(\theta)$.

Proof: The function $\psi(x) = (e^x - 1 - x)/x^2$ is increasing on $[0,\infty)$. Hence if θ is bounded by M, then

$$\begin{aligned} M^2\mathrm{E}\Big(e^{|t\varepsilon\theta(X)|/M} - 1 - \frac{|t\varepsilon\theta(X)|}{M}\Big) &= \mathrm{E}\psi\Big(\frac{t|\varepsilon\theta(X)|}{M}\Big)t^2\varepsilon^2\theta^2(X) \\ &\le \mathrm{E}\psi\big(t|\varepsilon|\big)t^2\varepsilon^2\theta^2(X) \le \mathrm{E}e^{t|\varepsilon|}\theta^2(X) \le \|r_t\|_\infty \mathrm{E}\theta^2(X), \end{aligned}$$

since $x^2\psi(x) \le e^x$ on $[0,\infty)$. It follows that $\big(M, 2\|\theta\|^2\|r_t\|_\infty\big)$ is a pair of Bernstein numbers for the variable $t\varepsilon\theta(X)$, and hence $\big(2M/t, 8\|\theta\|^2\|r_t\|_\infty/t^2\big)$ is a pair of Bernstein numbers for the variable $2\varepsilon\theta(X)$. Because $\big(M, e\|\theta\|^2\big)$ is a Bernstein pair for the variable $\theta(X)$ and $\theta^2 \le M|\theta|$ we have that $\big(M^2, e\|\theta\|^2M^2\big)$ is a Bernstein pair for the variable $\theta^2(X)$. Then the assertion follows from (iii) in Section 2 as $L\big((X,Y),\theta\big) = 2\varepsilon(\theta_0 - \theta)(X) + (\theta - \theta_0)^2(X)$ is the sum of two variables of this type.

The second assertion of the lemma is immediate. □

Corollary 4.2 *If the regression functions $\theta \in \Theta$ are bounded by a constant $M \ge 1$ and the error distribution has exponential tails, then, for any $\delta \in (0,1)$,*

$$\mathrm{E}\big\|\theta_{\hat{k}}(\mathbb{P}_S^0) - \theta_0\big\|^2 \le (1+2\delta)\inf_{k\in\mathcal{K}}\mathrm{E}\big\|\theta_k(\mathbb{P}_S^0) - \theta_0\big\|^2 + O\Big(\frac{1}{n}\Big)\log(1+\#\mathcal{K})\frac{M^2}{\delta}.$$

Proof: This is immediate from Theorem 2.3 (with $p = 1$) and the preceding lemma. □

Example 4.3 (adaptation to smoothness). To illustrate the strength of the method of cross validation we shall now use Theorem 3.2 to construct estimators that are adaptive to the full scale of Hölder spaces. Suppose that $\mathcal{X} = [0,1]$ and for each $(\alpha, M) \in (0,\infty)\times\mathbb{N}$ let $C_M^\alpha[0,1]$ be the set of functions $\theta\colon [0,1] \to \mathbb{R}$ that possess derivatives to the order the smallest integer $\underline{\alpha}$ strictly smaller than α which are bounded by M and with

the $\underline{\alpha}$th derivative satisfying $|\theta^{(\underline{\alpha})}(x) - \theta^{(\underline{\alpha})}(y)| \leq M|x-y|^{\alpha-\underline{\alpha}}$ for every $x, y \in [0,1]$. Assume that X possesses a density that is bounded away from zero and infinity. It is well known that, for each (α, M) (see e.g. Tsybakov (2004)), as $n \to \infty$, and certain constants C_α,

$$\inf_{\hat{\theta}} \sup_{\theta_0 \in C_M^\alpha[0,1]} \mathrm{E}_{\theta_0} \|\hat{\theta} - \theta_0\|^2 \asymp C_\alpha M^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)}.$$

The left side is the minimax risk for estimating θ_0 when θ_0 is known to belong to $C_M^\alpha[0,1]$, the infimum being over all estimators $\hat{\theta}$ based on a sample of n observations in the regression model $Y = \theta_0(X) + \varepsilon$. For each pair (α, M) let $\theta_{\alpha,M}(\mathbb{P}_S^0)$ be an estimator that is minimax up to a constant depending on α only and satisfies $\|\theta_{\alpha,M}(\mathbb{P}_0^S)\|_\infty \leq M$. We aim at choosing a pair $(\hat{\alpha}, \hat{M})$ that yields an estimator that is minimax (up to constants) for any (α, M).

Set $l_n = \log n$ and let $\mathcal{K} = \{(i/l_n, j) \colon i, j = 1, \ldots, n\}$. Then $\#\mathcal{K} \leq n^2$ and we may choose $\hat{k} = (\hat{\alpha}, \hat{M})$ from $\mathcal{K}$ by minimizing the penalized cross validated risk

$$(\alpha, M) \mapsto \mathrm{E}_S \int \big(y - \theta_{\alpha,k}(\mathbb{P}_S^0)(x)\big)^2 \, d\mathbb{P}_S^1(x,y) + \frac{\|\theta_{\alpha,k}(\mathbb{P}_S^0)\|_\infty^4 \vee 1}{n}.$$

By Lemma 4.1, $M(\theta) \lesssim \|\theta - \theta_0\|_\infty^2 \vee 1$ and $v(\theta) \leq R(\theta)\big(\|\theta - \theta_0\|_\infty^2 \vee 1\big)$. In view of Theorem 3.2 with $\delta = 1/4$ and $\lambda(\theta) = \|\theta\|_\infty^4 \vee 1$ there exists a constant C such that

$$\begin{aligned}\mathrm{E}_{\theta_0} \|\theta_{\hat{\alpha},\hat{M}}(\mathbb{P}_0^S) - \theta_0\|^2 \;\leq\; & 2 \inf_{(\alpha,j)\in\mathcal{K}} \Big[\mathrm{E}_{\theta_0} \|\theta_{\alpha,j}(\mathbb{P}_0^S) - \theta_0\|^2 + \frac{j^4}{n}\Big] \\ & + C(\log \#\mathcal{K})^2 \frac{1}{n} \sup_\theta \Big(\frac{\|\theta - \theta_0\|_\infty^2 \vee 1}{\|\theta\|_\infty^2 \vee 1}\Big)^2.\end{aligned}$$

Fix some $(\alpha, M) \in (0,\infty)^2$. As soon as n is sufficiently large that $\alpha < n/\log n$ and $M < n$, we have that there exists $(\alpha_n, j) \in \mathcal{K}$ with $|\alpha_n - \alpha| \leq l_n^{-1}$ and $|M - j| < 1$, and $C_M^\alpha[0,1] \subset C_j^{\alpha_n}[0,1]$. For any such (α_n, j), and every $\theta_0 \in C_M^\alpha[0,1]$,

$$\begin{aligned}&\mathrm{E}_{\theta_0} \|\theta_{\hat{\alpha},\hat{M}}(\mathbb{P}_0^S) - \theta_0\|^2 \\ &\quad \leq 2\Big[\mathrm{E}_{\theta_0} \|\theta_{\alpha_n,j}(\mathbb{P}_0^S) - \theta_0\|^2 + \frac{j^4}{n}\Big] + C(\log \#\mathcal{K})^2 \frac{1}{n}\big(1 + \|\theta_0\|_\infty\big)^4 \\ &\quad \lesssim j^{2/(2\alpha+1)} n^{-2\alpha_n/(2\alpha_n+1)} + M^4/n + M^4(\log n)^2/n.\end{aligned}$$

It follows that, for every (α, M),

$$\limsup_{n\to\infty} M^{-2/(2\alpha+1)} n^{2\alpha/(2\alpha+1)} \sup_{\theta_0 \in C_M^\alpha[0,1]} \mathrm{E}_{\theta_0} \|\theta_{\hat{\alpha},\hat{M}}(\mathbb{P}_0^S) - \theta_0\|^2 \lesssim D_\alpha < \infty.$$

Thus the estimator $\theta_{\hat{\alpha},\hat{M}}(\mathbb{P}_S^0)$ is asymptotically minimax on every Hölder ball up a to constant depending only on α.

As a purely asymptotic result the existence of estimators with this property could also be derived without using penalties, but by allowing at "time" n only estimators that are

bounded by a constant M_n, with M_n increasing slowly (logarithmically) to infinity. The remainder term in Theorem 2.3 would then be $O(M_n^2/n)$. However, this approach seems to be of "asymptopia" character (a mathematically correct limit result, but practically useless), because one would have to "wait" a long time before a large θ_0 (larger than M_n for reasonable n) would even be within the scope of the estimators.

5 Least absolute deviation regression

Consider the regression model $Y = \theta_0(X) + \varepsilon$ of Example 1.1, with error ε with zero conditional median: $\mathrm{P}(\varepsilon \le 0 | X) = \mathrm{P}(\varepsilon \ge 0 | X) = 1/2$.

The mean absolute deviation criterion, centered at its minimum, satisfies

$$\big|L\big((X,Y),\theta\big)\big| = \Big|\big|Y-\theta(X)\big| - \big|Y-\theta_0(X)\big|\Big| \le |\theta-\theta_0|(X).$$

This shows that the loss function is bounded as soon as the regression functions are bounded, irrespective of the error distribution.

Assume that the error distribution has finite absolute moment and is smooth enough at its median 0 in order that, for $\mu \in \mathbb{R}$,

$$\mu^2 \wedge |\mu| \lesssim \mathrm{E}|\varepsilon-\mu| - \mathrm{E}|\varepsilon| \lesssim \mu^2 \wedge |\mu|, \tag{5.1}$$

where the constants in the inequalities may depend on the error distribution. The absolute value $|\mu|$ is not necessary if μ is restricted to a compact interval around the origin, but cannot be missed in general, as $|\varepsilon-\mu|$ grows sub-linearly as $\mu \to \infty$. Under this condition the risk is equivalent to a mixed L_1-L_2 distance

$$R(\theta) = \mathrm{E}\big|\varepsilon - (\theta-\theta_0)(X)\big| - \mathrm{E}|\varepsilon| \asymp P\big((\theta-\theta_0)^2 \wedge |\theta-\theta_0|\big). \tag{5.2}$$

Lemma 5.1 *If the regression functions $\theta \in \Theta$ are bounded and the error distribution satisfies (5.1), then $\big(M(\theta), v(\theta)\big) = \big(\|\theta-\theta_0\|_\infty, 1.5P(\theta-\theta_0)^2\big)$ are Bernstein pairs for the functions $x \mapsto L(x,\theta)$. Furthermore $v(\theta) \lesssim \big(\|\theta-\theta_0\|_\infty \vee 1\big)R(\theta)$.*

Proof: If $\big(M(\theta), v(\theta)\big)$ is a Bernstein pair for the variable $(\theta-\theta_0)(X)$, then so it is for the variable $L\big((X,Y),\theta\big) - L\big((X,Y),\theta_0\big)$. In particular, we may use the Bernstein pair $\big(\|\theta-\theta_0\|_\infty, 1.5P(\theta-\theta_0)^2\big)$.

The second assertion follows with the help of (5.1) in view of the identity $x^2 = (x \vee 1)(x^2 \wedge x)$ for every $x \ge 0$ applied to $x = |\theta-\theta_0|(X)$. □

Corollary 5.2 *If the regression functions $\theta \in \Theta$ are bounded by a constant $M \ge 1$ and the error distribution satisfies (5.1), then, for any $\delta \in (0,1)$,*

$$\mathrm{E}R\big(\theta_{\hat{k}}(\mathbb{P}_S^0)\big) \le (1+2\delta) \inf_{k \in \mathcal{K}} \mathrm{E}R\big(\theta_k(\mathbb{P}_S^0)\big) + O\Big(\frac{1}{n}\Big) \log(1+\#\mathcal{K})\, \frac{M}{\delta}.$$

Proof: We apply Theorem 2.3 with $p = 1$. □

The risk $R(\theta)$ is the prediction error relative to the absolute deviation. Under the assumption of boundedness of the regression functions, it is up to constants the square L_2-distance $\|\theta - \theta_0\|^2$ as in Section 4, in view of (5.2).

6 Classification

Consider the classification problem of Example 1.2 with loss function

$$L\big((x, y), \theta\big) = 1_{y \neq \theta(x)} - 1_{y \neq \theta_0(x)},$$

for θ_0 the Bayes classifier $\theta_0 = 1_{\eta_0 \geq 1/2}$. The corresponding risk function is the probability of misclassification, centered at its minimum value:

$$R(\theta) = \mathrm{P}\big(Y \neq \theta(X)\big) - \mathrm{P}\big(Y \neq \theta_0(X)\big).$$

A natural distance in this problem is the L_1-distance

$$d(\theta_1, \theta_2) = \mathrm{E}\big|1_{Y \neq \theta_1(X)} - 1_{Y \neq \theta_2(X)}\big| = \mathrm{P}\big(\theta_1(X) \neq \theta_2(X)\big).$$

Tsybakov's condition (Mammen and Tsybakov (1999), Tsybakov (2004)) relates this distance to the risk. It requires that, for some $\gamma \geq 1$ and positive constant D,

$$R(\theta) - R(\theta_0) \geq D\, d(\theta, \theta_0)^\gamma. \tag{6.1}$$

The condition can be viewed as measuring the probability that an input X gives rise to a Bayes classifier $\eta_0(X)$ that is close to the decision boundary $1/2$. Proposition 1 in Tsybakov (2004) shows that the condition is satisfied with $\gamma = 1 + \alpha^{-1}$ if $\mathrm{P}\big(|\eta_0(X) - 1/2| \leq t\big) \lesssim t^\alpha$ for $t > 0$.

If η_0 is bounded away from $1/2$, then (6.1) is satisfied with $\gamma = 1$ (limiting case $\alpha = \infty$), which is the most favorable situation for estimating θ. In this case also the remainder in the following oracle inequality is smallest: order $O(1/n)$ times the logarithmic complexity of the set of estimators. For $\gamma > 1$ both the typical rate of "learning", the decrease of $\mathrm{E}R\big(\theta(\mathbb{P})\big) - R(\theta_0)$ for an appropriate estimator, and the remainder in the oracle inequality are bigger. For the situation considered in Theorem 1 of Tsybakov (2004) such a typical rate of learning is $n^{-\gamma/(2\gamma - 1 + \rho)}$ for $\rho \in (0, 1)$ a parameter measuring the complexity of the set of classification functions. The remainder in the oracle inequality in the following theorem is smaller for any such ρ.

It may be noted that condition (6.1) is satisfied for any $\gamma > \gamma_0$ if it is satisfied for $\gamma = \gamma_0$. Therefore we can alway apply the oracle inequality with $\gamma = \infty$, in which case the remainder is $O(n^{-1/2})$ and the choice $\delta = 0$ is eligible.

Lemma 6.1 *The pairs $\big(M(\theta), v(\theta)\big) = \big(1, 1.5 d(\theta, \theta_0)\big)$ are Bernstein pairs for the functions $x \mapsto L(x, \theta)$. Furthermore, if (6.1) is satisfied, then $v(\theta) \leq 1.5 D^{-1/\gamma} R(\theta)^{1/\gamma}$.*

Proof: The loss function has range $\{-1, 0, 1\}$ and hence is bounded by 1, so that 1 together with e times the variance of the loss function forms a Bernstein pair. The variance is bounded by the second moment, which is $d(\theta, \theta_0)$. The second assertion of the lemma is immediate. □

Corollary 6.2 *If (6.1) is satisfied for some $\gamma \geq 1$, then for any $\delta \in (0, 1)$,*

$$\begin{aligned}\mathrm{E}R\big(\theta_{\hat{k}}(\mathbb{P}_S^0)\big) \;&\leq (1+2\delta)\inf_{k\in\mathcal{K}} \mathrm{E}R\big(\theta_k(\mathbb{P}_S^0)\big) \\ &\quad +(1+\delta)\mathrm{E}\Big(\frac{16}{(n^1)^{\gamma/(2\gamma-1)}}\Big)\log(1+\#\mathcal{K})\Big[1+\Big(\frac{1+\delta}{\delta D}\Big)^{1/(2\gamma-1)}e\Big].\end{aligned}$$

Proof: We apply Theorem 2.3 with $2 - p = 1/\gamma$, so that $1/p = \gamma/(2\gamma - 1)$. □

7 Multivariate mean

Consider the problem of estimating the mean vector $\theta_0 \in \mathbb{R}^D$ of a sample from the distribution of $X = \theta_0 + \varepsilon$, for ε a D-dimensional standard normal vector (see Example 1.3), relative to the (centered) loss function

$$L(X, \theta) = \|X - \theta\|^2 - \|X - \theta_0\|^2 = 2(\theta_0 - \theta)^T\varepsilon + \|\theta - \theta_0\|^2.$$

The corresponding risk function is the square Euclidean distance $R(\theta) = \|\theta - \theta_0\|^2$.

If θ ranges freely over $\mathbb{R}^D$, then the loss function is unbounded, and the risk estimator obtained from cross validation, even though unbiased, suffers from a large variance. This motivates the introduction of a penalty. Consider the criterion

$$k \mapsto \mathrm{E}_S \int \big\|x - \theta_k(\mathbb{P}_S^0)\big\|^2 \, d\mathbb{P}_S^1(x) + \frac{\big\|\theta_k(\mathbb{P}_S^0)\big\|^2 + 1}{n}. \tag{7.1}$$

Lemma 7.1 *The pairs $\big(M(\theta), v(\theta)\big) = \big(\|\theta - \theta_0\|, 4e^2\|\theta - \theta_0\|^2\big)$ are Bernstein pairs for the functions $x \mapsto L(x, \theta) - \|\theta - \theta_0\|^2$.*

Proof: The variable $L(X, \theta) - \|\theta - \theta_0\|^2 = 2(\theta - \theta_0)^T\varepsilon$ is distributed as $2\|\theta - \theta_0\|Z$ for a univariate standard normal variable Z, and

$$M^2\mathrm{E}\Big(e^{2|Z|\|\theta-\theta_0\|/M} - 1 - \frac{2|Z|\|\theta - \theta_0\|}{M}\Big) = \sum_{k\geq 2}\frac{\mathrm{E}|2Z|^2}{k!}\Big(\frac{\|\theta-\theta_0\|}{M}\Big)^{k-2}\|\theta - \theta_0\|^2$$

is bounded above by $\|\theta - \theta_0\|^2\mathrm{E}e^{2|Z|} \leq \|\theta - \theta_0\|^2 2e^2$, for $M \geq \|\theta - \theta_0\|$. The result follows. □

Corollary 7.2 *For any $\delta \in (0,1)$, the minimizer $\hat{k}$ of (7.1) satisfies, for a universal constant C,*

$$\mathrm{E}\big\|\theta_{\hat{k}}(\mathbb{P}_S^0) - \theta_0\big\|^2 \le (1+2\delta) \inf_{k\in\mathcal{K}}\Big[\mathrm{E}\big\|\theta_k(\mathbb{P}_S^0) - \theta_0\big\|^2 + \frac{\big\|\theta_k(\mathbb{P}_S^0)\big\|^2}{n}\Big] + C\,\mathrm{E}\frac{1}{n^1}\frac{1}{\delta^3}\big[\log(1+\#\mathcal{K})\big]^2\big(\|\theta_0\|^2+1\big).$$

Proof: We apply Theorem 3.2 with the penalty $\lambda(k,\theta) = \|\theta\|^2 + 1$. Because $L(x,\theta) = 2(\theta-\theta_0)^T\varepsilon + \|\theta-\theta_0\|^2$ is up to a constant equal to $2(\theta-\theta_0)^T\varepsilon$ and the empirical process maps constants into 0, we may take the Bernstein pairs $\big(M(\theta), v(\theta)\big)$ in the application of Theorem 3.2 as in the preceding lemma. Then

$$\sup_\theta\Big(\frac{M(\theta)}{\sqrt{\lambda(\theta)}}\Big)^2 \le \sup_\theta \frac{\|\theta-\theta_0\|^2}{\|\theta\|^2+1} \le 2 + 2\|\theta_0\|^2,$$
$$\sup_\theta \frac{v(\theta)}{R(\theta)\sqrt{\lambda(\theta)}} \le \sup_\theta \frac{4e^2\|\theta-\theta_0\|^2}{\|\theta-\theta_0\|^2(\|\theta\|^2+1)^{1/2}} \le 4e^2.$$

Therefore the assertion of Theorem 3.2 simplifies to the present inequality. □

Under the assumption that at least one of the estimators $\theta_k(\mathbb{P}_S^0)$ is consistent for θ_0, the penalty $\big\|\theta_k(\mathbb{P}_S^0)\big\|^2/n$ inside the infimum over $k \in \mathcal{K}$ in the corollary will contribute of the order $O_P(1/n)$, which is smaller than the remainder. Somewhat remarkably, the bound of the corollary is dimensionless: the dimension D, which may be very large, enters only through the risks of the estimators $\theta_k(\mathbb{P}_S^0)$ and the norm $\|\theta_0\|^2$, not through the cross validation.

The estimators $\theta_k(\mathbb{P}_S^0)$ could be constructed in many ways. For instance, each $\theta_k(\mathbb{P}_S^0)$ could be a penalized least squares estimator

$$\theta_k(\mathbb{P}_S^0) = \operatorname*{argmin}_{\theta\in\mathbb{R}^D} \int \|x-\theta\|^2\, d\mathbb{P}_S^0(x) + \mu_k \frac{\|\theta\|_r^r}{n},$$

with the smoothing parameter μ_k, which controls the influence of the penalty, ranging over a grid in an interval $(0,\mu)$. The values $r=1$ and $r=2$ correspond to the LASSO and ridge regression estimator, respectively. Alternatively, we could hypothesize that the mean vector θ is sparse and construct an estimator under the assumption that at most k coordinates θ_i are nonzero. We can cross-validate over a set of estimators containing an estimator appropriate for each subset $I \subset \{1,\ldots,D\}$ of nonzero coordinates with $\#I \le K$ for a given constant K, as this gives a set of $\#\mathcal{K} \le D^K$ estimators. However, the preceding theorem does not allow a useful conclusion for cross validation over all subsets $I \subset \{1,\ldots,D\}$, as $\#\mathcal{K}$ would be 2^D in that case, yielding a remainder term of the order D^2/n. This shows the limitation of the theorem: because it applies to arbitrary estimators $\theta_k(\mathbb{P}_S^0)$ without regard of relationships between the estimators, the remainder can be pessimistic, even if it is logarithmic. Minimum penalized contrast estimators can adapt to all subsets, as shown in Birgé and Massart (2001). Such a result is also obtainable by a double cross validation, along the lines of van der Laan et al. (2006).

8 Proofs

In this section we gather technical proofs.

Lemma 8.1 *Let $X_1, \ldots, X_m$ be arbitrary random variables such that $\mathrm{P}(X_i > x) \leq K_i e^{-C_i x^p}$ for every $x > 0$ and for given positive constants K_i and C_i and p. Then, with $C = \min_{1 \leq i \leq m} C_i$ and D_p a constant depending only on p (with $D_p = 0$ if $p \geq 1$),*

$$\mathrm{E} \max_{1 \leq i \leq m} X_i \leq \Big(\frac{2}{C} \log\Big(1 + \sum_{i=1}^m \frac{CK_i}{C_i} + D_p\Big)\Big)^{1/p}.$$

Proof: We may assume that the variables X_i are nonnegative; otherwise we replace them by the variables X_i^+.

For $p \geq 1$ the function $x \mapsto \psi(x) = e^{x^p} - 1$ is nonnegative, convex and nondecreasing on $[0, \infty)$. Therefore, by Jensen's inequality,

$$\psi\Big(d^{1/p} \mathrm{E} \max_i X_i\Big) \leq \mathrm{E}\psi\Big(d^{1/p} \max_i X_i\Big) = \mathrm{E} \max_i \psi\Big(d^{1/p} X_i\Big) \leq \sum_{i=1}^m \mathrm{E}\psi\Big(d^{1/p} X_i\Big).$$

We can compute the expectations in the right side as

$$\mathrm{E}\psi\Big(d^{1/p} X_i\Big) = \mathrm{E} \int_0^{d^{1/p} X_i} e^{x^p} p x^{p-1}\, dx = \int_0^\infty \mathrm{P}(d^{1/p} X_i > x) e^{x^p} p x^{p-1}\, dx,$$

by Fubini's theorem. We can now insert the upper tail bound, and calculate the resulting integral as $dK_i/(C_i - d)$, provided that $d < C_i$. For $d = \frac{1}{2} \min_i C_i$ we have that $C_i - d \geq \frac{1}{2} C_i$ and hence $dK_i/(C_i - d) \leq CK_i/C_i$. We substitute this bound in the preceding display, and next apply the function $\psi^{-1}(m) = \big(\log(1+m)\big)^{1/p}$ left and right to the inequality.

For $0 < p \leq 1$ the function $x \mapsto \psi(x) = e^{x^p} - 1$ is convex only on the interval $[e_p, \infty)$, for $e_p = (p^{-1} - 1)^{1/p}$, and the preceding argument must be adapted. We define a function $\tilde{\psi}$ by $\tilde{\psi}(x) = \psi(x)$ for $x > e_p$ and $\tilde{\psi}$ constant and continuous on the interval $[0, e_p]$. Then $\tilde{\psi}$ is convex, satisfies $\tilde{\psi} \leq \psi + E_p$ for $E_p = \psi\big(e_p\big)$ on $[0, \infty)$, and is strictly increasing on $[e_p, \infty)$ with the same inverse as ψ. Applying the preceding argument with $\tilde{\psi}$ instead of ψ gives that $\tilde{\psi}\big(d^{1/p} \mathrm{E} \max X_i\big)$ is bounded by $\sum_{i=1}^m \mathrm{E}\psi(d^{1/p} X_i) + E_p$, where $\mathrm{E}\psi(d^{1/p} X_i)$ is bounded by $dK_i/(C_i - d)$, as before. Hence $d^{1/p} \mathrm{E} \max X_i$ is bounded by e_p or is bounded by $\psi^{-1}\big(\sum_i (CK_i/C_i) + E_p\big)$. This implies the result, for a sufficiently large constant D_p. □

Lemma 8.2 *Let $\mathbb{G}$ be the empirical process of an i.i.d. sample of size n from the distribution P and let $\lambda \colon \mathcal{F} \to (0, \infty)$ be arbitrary. Then, for any Bernstein pairs $\big(M(f), v(f)\big)$*

and for any $\delta > 0$, $0 < p \leq 2$ *and* $0 < q \leq 1$,

$$\begin{aligned}
&\mathrm{E} \max_{f \in \mathcal{F}} \bigl(\mathbb{G} f - \lambda(f) \bigr) \\
&\quad \leq \frac{1}{n^{1/(2q)}} \Bigl[\log \Bigl(1 + \sum_{f \in \mathcal{F}} e^{-C_q \sqrt{n} \lambda(f)/(4M(f))} + D_q \Bigr) \Bigr]^{1/q} \Bigl(\max_{f \in \mathcal{F}} \frac{8M(f)}{C_q \lambda(f)^{1-q}} \Bigr)^{1/q} \\
&\qquad\qquad + \Bigl[\log \Bigl(1 + \sum_{f \in \mathcal{F}} e^{-C_p \lambda(f)^2/(4v(f))} + D_p \Bigr) \Bigr]^{1/p} \max_{f \in \mathcal{F}} \Bigl(\frac{8v(f)}{C_p \lambda(f)^{2-p}} \Bigr)^{1/p}.
\end{aligned}$$

Here $C_p > 0$ *and* $D_p \geq 0$ *are constants with* $C_p = 1$ *and* $D_p = 0$ *for* $p \geq 1$. *The same upper bound is valid for* $\mathrm{E} \max_{f \in \mathcal{F}} \bigl(\mathbb{G}(-f) - \lambda(f) \bigr)$.

Proof: By Bernstein's inequality (e.g. van der Vaart and Wellner (1996), Lemma 2.2.11), for every $x > 0$,

$$\mathrm{P} \bigl(\mathbb{G} f - \lambda(f) > x \bigr) \leq e^{-\frac{1}{2} \frac{(x+\lambda(f))^2}{v(f) + (x+\lambda(f)) M(f)/\sqrt{n}}}.$$

The quotient in the exponent can be bounded further by using the inequalities, with $b = M/\sqrt{n}$ and $r = v/b - \lambda$,

$$\frac{(x+\lambda)^2}{v + (x+\lambda)b} \geq \begin{cases} \frac{(x+\lambda)^2}{2v} \geq \frac{(x+\lambda)^p \lambda^{2-p}}{2v} \geq C_p \frac{x^p \lambda^{2-p} + \lambda^2}{2v}, & \text{if } x \leq r, \\ \frac{x+\lambda}{2b} \geq \frac{(x+\lambda)^q \lambda^{1-q}}{2b} \geq C_q \frac{x^q \lambda^{1-q} + \lambda}{2b}, & \text{if } x \geq r. \end{cases}$$

Here C_p is the constant in the inequality $(x + \lambda)^p \geq C_p (x^p + \lambda^p)$, which can be taken equal to 1 for $p \geq 1$ and equal to 2^{p-1} for $0 < p \leq 1$. It follows that, for *all* $x > 0$,

$$\begin{aligned}
\mathrm{P} \bigl((\mathbb{G} f - \lambda(f)) 1_{(\mathbb{G} f - \lambda(f)) \leq r} > x \bigr) &\leq e^{-C_p \frac{x^p \lambda^{2-p} + \lambda^2}{4v}}, \\
\mathrm{P} \bigl((\mathbb{G} f - \lambda(f)) 1_{(\mathbb{G} f - \lambda(f)) \geq r} > x \bigr) &\leq e^{-C_q \frac{x^q \lambda^{1-q} + \lambda}{4b}}.
\end{aligned}$$

Two applications of Lemma 8.1, with the constants taken equal to $K_f = e^{-C_p \lambda^2/(4v)}$ and $C = C_f = C_p \lambda^{2-p}/(4v)$, and $K_f = e^{-C_q \lambda/(4b)}$ and $C = C_f = C_q \lambda^{1-q}/(4b)$, respectively, yield that, with $Y_{\leq r} = Y 1_{Y \leq r}$ and $Y_{> r} = Y 1_{Y > r}$,

$$\begin{aligned}
\mathrm{E} \max_f (\mathbb{G} f - \lambda(f))_{\leq r} &\leq \max_f \Bigl(\frac{8v(f)}{C_p \lambda(f)^{2-p}} \Bigr)^{1/p} \Bigl[\log \Bigl(1 + \sum_f e^{-C_p \lambda^2/(4v)} + D_p \Bigr) \Bigr]^{1/p}, \\
\mathrm{E} \max_f (\mathbb{G} f - \lambda(f))_{> r} &\leq \max_f \Bigl(\frac{8b(f)}{C_q \lambda(f)^{1-q}} \Bigr)^{1/q} \Bigl[\log \Bigl(1 + \sum_f e^{-C_q \lambda/(4b)} + D_q \Bigr) \Bigr]^{1/q}.
\end{aligned}$$

Adding these equations, substituting $b(f) = M(f)/\sqrt{n}$ and rearranging gives the result. □

Proof:of Lemma 3.1. We apply the preceding lemma with the numbers $\lambda(f)$ replaced by the numbers $\delta \sqrt{n} (Pf + \lambda(f)/n)$. We can bound the resulting denominator $\bigl(\delta \sqrt{n} (Pf +$

$\lambda(f)/n)\big)^{1-q}$ in the first term maximum on the right below by $\big(\delta\lambda(f)/\sqrt{n}\big)^{1-q}$, and the denominator $\big(\delta\sqrt{n}(Pf+\lambda(f)/n)\big)^{2-p}$ in the second maximum on the right below by $(\delta\sqrt{n})^{2-p}(Pf)^{(2-p)(1-s)}(\lambda(f)/n)^{(2-p)s}$, for $s=(1-p)/(2-p)$, so that $(2-p)s=1-p$ and $(2-p)(1-s)=1$. □

8.1 Bernstein numbers

In this subsection we prove properties (i)-(iv) of Bernstein numbers as given in Section 2. For (i) we note that

$$M^2P\Big(e^{|f|/M}-1-\frac{|f|}{M}\Big)=M^2\sum_{k\geq 2}P\frac{|f|^k}{k!M^k}\leq Pf^2\sum_{k\geq 2}\frac{\|f\|_\infty^{k-2}}{k!M^{k-2}}\leq Pf^2\sum_{k\geq 2}\frac{1}{k!},$$

if $\|f\|_\infty\leq M$. The series is equal to $e-2$. Property (ii) is clear from the definition. For (iii) we note that, because the function $\psi(x)=e^x-1-x$ is convex and increasing on $[0,\infty)$,

$$M^2P\psi\Big(\frac{|f+g|}{M}\Big)\leq\tfrac{1}{2}M^2P\psi\Big(\frac{2|f|}{M}\Big)+\tfrac{1}{2}M^2P\psi\Big(\frac{2|g|}{M}\Big).$$

This is bounded by $v(f)+v(g)$ if $M\geq 2M(f)$ and $M\geq 2M(g)$, as the function $M\mapsto\psi(x/M)$ is decreasing for every $x\geq 0$. Property (iv) is proved similarly.

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest, 1973.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans. Automatic Control*, AC-19:716–723, 1974. ISSN 0018-9286.

Donald W. K. Andrews. Asymptotic optimality of generalized C_L, cross-validation, and generalized cross-validation in regression with heteroskedastic errors. *J. Econometrics*, 47(2-3):359–377, 1991. ISSN 0304-4076.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probab. Theory Related Fields*, 113(3):301–413, 1999. ISSN 0178-8051.

Andrew R. Barron and Thomas M. Cover. Minimum complexity density estimation. *IEEE Trans. Inform. Theory*, 37(4):1034–1054, 1991. ISSN 0018-9448.

L. Birgé. Statistical estimation with model selection. Technical report, Brouwer lecture, 2006.

Lucien Birgé and Pascal Massart. Gaussian model selection. *J. Eur. Math. Soc. (JEMS)*, 3(3):203–268, 2001. ISSN 1435-9855.

Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: a survey of some recent advances. *ESAIM Probab. Stat.*, 9:323–375 (electronic), 2005. ISSN 1292-8100.

F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for regression learning. Technical report, 2004.

Simon L. Davies, Andrew A. Neath, and Joseph E. Cavanaugh. Cross validation model selection criteria for linear regression based on the Kullback-Leibler discrepancy. *Stat. Methodol.*, 2(4):249–266, 2005. ISSN 1572-3127.

Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001. ISBN 0-387-95117-2.

Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Stat. Methodol.*, 2(2): 131–154, 2005. ISSN 1572-3127.

Bradley Efron. The estimation of prediction error: covariance penalties and cross-validation. *J. Amer. Statist. Assoc.*, 99(467):619–642, 2004. ISSN 0162-1459.

Edward I. George. The variable selection problem. *J. Amer. Statist. Assoc.*, 95(452): 1304–1308, 2000. ISSN 0162-1459.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 0-387-95441-4.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Trans. Inform. Theory*, 47(5):1902–1914, 2001. ISSN 0018-9448.

Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.*, 15(3):958–975, 1987. ISSN 0090-5364.

Gábor Lugosi and Andrew B. Nobel. Adaptive model selection using empirical complexities. *Ann. Statist.*, 27(6):1830–1864, 1999. ISSN 0090-5364.

Colin L. Mallows. Some comments on C_p. *Technometrics*, 15:661–671, 1973.

Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 1999. ISSN 0090-5364.

Pascal Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse Math. (6)*, 9(2):245–303, 2000. ISSN 0240-2963.

Arkadi Nemirovski. Topics in non-parametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1998)*, volume 1738 of *Lecture Notes in Math.*, pages 85–277. Springer, Berlin, 2000.

M. Stone. Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, 36:111–147, 1974. ISSN 0035-9246.

M. Stone. Corrigendum: "Cross-validatory choice and assessment of statistical predictions" (J. Roy. Statist. Soc. Ser. B **36** (1974), 111–147). *J. Roy. Statist. Soc. Ser. B*, 38 (1):102, 1976. ISSN 0035-9246.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004. ISSN 0090-5364.

Sara van de Geer. Least squares estimation with complexity penalties. *Math. Methods Statist.*, 10(3):355–374, 2001. ISSN 1066-5307.

M.J. van der Laan, S. Dudoit, and A.W. van der Vaart. The cross-validated epsilon net estimator. *preprint*, 2006.

Aad W. van der Vaart and Jon A. Wellner. *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York, 1996. ISBN 0-387-94640-3.

Vladimir N. Vapnik. *Statistical learning theory*. Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons Inc., New York, 1998. ISBN 0-471-03003-1.

Marten Wegkamp. Model selection in nonparametric regression. *Ann. Statist.*, 31(1): 252–273, 2003. ISSN 0090-5364.

Yuhong Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000. ISSN 0090-5364.

Ping Zhang. Model selection via multifold cross validation. *Ann. Statist.*, 21(1):299–313, 1993. ISSN 0090-5364.

Aad van der Vaart
Department of Mathematics
Vrije Universiteit Amsterdam
aad@cs.vu.nl

Mark van der Laan
Sandrine Dudoit
Division of Biostatistics
University of California, Berkeley
laan@stat.berkeley.edu