A Stochastic Gradient Descent Algorithm for Structural Risk Minimisation

Joel Ratsaby

Department of Computer Science, University College London, Gower Street,
London WC1E 6BT, U.K.

Abstract

Structural risk minimisation (SRM) is a general complexity regularization method which automatically selects the model complexity that approximately minimises the misclassification error probability of the empirical risk minimiser. It does so by adding a complexity penalty term $\epsilon(m,k)$ to the empirical risk of the candidate hypotheses and then for any fixed sample size m it minimises the sum with respect to the model complexity variable k.

When learning multicategory classification there are M subsamples m_i, corresponding to the M pattern classes with *a priori* probabilities p_i, $1 \leq i \leq M$. Using the usual representation for a multi-category classifier as M individual boolean classifiers, the penalty becomes $\sum_{i=1}^{M} p_i \epsilon(m_i, k_i)$. If the m_i are given then the standard SRM trivially applies here by minimizing the penalised empirical risk with respect to k_i, $1, \ldots, M$.

However, in situations where the total sample size $\sum_{i=1}^{M} m_i$ needs to be minimal one needs to also minimize the penalised empirical risk with respect to the variables m_i, $i = 1, \ldots, M$. The obvious problem is that the empirical risk can only be defined after the subsamples (and hence their sizes) are given (known).

Utilising an on-line stochastic gradient descent approach, this paper overcomes this difficulty and introduces a sample-querying algorithm which extends the standard SRM principle. It minimises the penalised empirical risk not only with respect to the k_i, as the standard SRM does, but also with respect to the m_i, $i = 1, \ldots, M$.

The challenge here is in defining a stochastic empirical criterion which when minimised yields a sequence of subsample-size vectors which asymptotically achieve the Bayes' optimal error convergence rate.

1 Introduction

Consider the general problem of learning classification with M pattern classes each with a class conditional probability density $f_i(x)$, $1 \leq i \leq M$, $x \in \mathbb{R}^d$,

Proc. of the 14th Int'l Conf. On Alg. Learn. Theory, Springer LNAI (2003) Vol.2842 ,

and *a priori* probabilities p_i, $1 \le i \le M$. The functions $f_i(x)$, $1 \le i \le M$, are assumed to be unknown while the p_i are assumed to be known or unknown depending on the particular setting. The learner observes randomly drawn i.i.d. examples each consisting of a pair of a feature vector $x \in \mathbb{R}^d$ and a label $y \in \{1, 2, \ldots, M\}$, which are obtained by first drawing y from $\{1, \ldots, M\}$ according to a discrete probability distribution $\{p_1, \ldots, p_M\}$ and then drawing x according to the selected probability density $f_y(x)$.

Denoting by $c(x)$ a classifier which represents a mapping $c : \mathbb{R}^d \to \{1, 2, \ldots, M\}$ then the *misclassification error* of c is defined as the probability of misclassification of a randomly drawn x with respect to the underlying mixture probability density function $f(x) = \sum_{i=1}^{M} p_i f_i(x)$. This misclassification error is commonly represented as the expected 0/1-loss, or simply as the *loss*, $L(c) = \mathrm{E}1_{\{c(x) \neq y(x)\}}$, of c where expectation is taken with respect to $f(x)$ and $y(x)$ denotes the true label (or class origin) of the feature vector x. In general $y(x)$ is a random variable depending on x and only in the case of $f_i(x)$ having non-overlapping probability 1 supports then $y(x)$ is a deterministic function [1] . The aim is to learn, based on a finite randomly drawn labelled sample, the optimal classifier known as the Bayes classifier which by definition has minimum loss. In this paper we pose the following question:

Question: Can the learning accuracy be improved if labeled examples are independently randomly drawn according to the underlying class conditional probability distributions but the pattern classes, i.e., the example labels, are chosen *not* necessarily according to their *a priori* probabilities ?

We answer this in the affirmative by showing that there exists a tuning of the subsample proportions which minimizes a loss criterion. The tuning is relative to the intrinsic complexity of the Bayes-classifier.

Before continuing let us introduce some notation. We write *const* to denote absolute constants or constants which do not depend on other variables in the mathematical expression. We denote by $\{(x_j, y_j)\}_{j=1}^{\overline{m}}$ an i.i.d. sample of labelled examples where $\overline{m}$ denotes the total sample size, y_j, $1 \le j \le \overline{m}$, are drawn i.i.d. and taking the integer value 'i' with probability p_i, $1 \le i \le M$, while the corresponding x_j are drawn according to the class conditional probability density $f_{y_j}(x)$. Denote by m_i the number of examples having a y-value of 'i'. Denote by $m = [m_1, \ldots, m_M]$ the sample size vector and let $\|m\| = \sum_{i=1}^{M} m_i \equiv \overline{m}$. The notation $\mathrm{argmin}_{k \in A} g(k)$ for a set A means the subset (of possibly more than one element) whose elements have the minimum value of g over A. A slight abuse of notation will be made by using it for countable sets where the notation means the subset of elements k such

[1] According to the probabilistic data-generation model mentioned above, only regions in probability 1 support of the mixture distribution $f(x)$ have a well-defined class membership.

that [2] $g(k) = \inf_{k'} g(k')$. The loss $L(c)$ is expressed in terms of the class-conditional losses, $L_i(c)$, as $L(c) = \sum_{i=1}^{M} p_i L_i(c)$ where $L_i(c) = \mathrm{E}_i 1_{\{c(x) \neq i\}}$, and E_i is the expectation with respect to the density $f_i(x)$. The empirical counterparts of the loss and conditional loss are $L_m(c) = \sum_{i=1}^{M} p_i L_{i,m_i}(c)$ where $L_{i,m_i}(c) = \frac{1}{m_i} \sum_{j: y_j = i} 1_{\{c(x_j) \neq i\}}$ where throughout the paper we assume the *a priori* probabilities are known to the learner (see Assumption 1 below).

2 Structural Risk Minimisation

The loss $L(c)$ depends on the unknown underlying probability distributions hence realistically for a learning algorithm to work it needs to use only an estimate of $L(c)$. For a finite class $\mathcal{C}$ of classifiers the empirical loss $L_m(c)$ is a consistent estimator of $L(c)$ uniformly for all $c \in \mathcal{C}$ hence provided that the sample size $\overline{m}$ is sufficiently large, an algorithm that minimises $L_m(c)$ over C will yield a classifier $\hat{c}$ whose loss $L(\hat{c})$ is an arbitrarily good approximation of the true minimum Bayes loss, denoted here as L^*, provided that the optimal Bayes classifier is contained in $\mathcal{C}$. The Vapnik-Chervonenkis theory [Vapnik, 1982] characterises the condition for such uniform estimation over an *infinite* class $\mathcal{C}$ of classifiers. The condition basically states that the class needs to have a finite complexity or richness which is known as the *Vapnik-Chervonenkis dimension* and is defined as follows: for a class H of functions from a set X to $\{0, 1\}$ and a set $S = \{x_1, \ldots, x_l\}$ of l points in X, denote by $H_{|S} = \{[h(x_1), \ldots, h(x_l)] : h \in H\}$. Then the Vapnik-Chervonenkis dimension of H denoted by $VC(H)$ is the largest l such that the cardinality $\left|H_{|S}\right| = 2^l$. The method known as *empirical risk minimisation* represents a general learning approach which for learning classification minimises the 0/1-empirical loss and provided that the hypothesis class has a finite VC dimension then the method yields a classifier $\hat{c}$ with an asymptotically arbitrarily-close loss to the minimum L^*.

As is often the case in real learning algorithms, the hypothesis class can be rich and may practically have an infinite VC-dimension, for instance, the class of all two layer neural networks with a variable number of hidden nodes. The method of *Structural Risk Minimisation* (SRM) was introduced by Vapnik [1982] in order to learn such classes via empirical risk minimisation.

For the purpose of reviewing existing results we limit our discussion for the remainder of this section to the case of two-category classification thus we use m and k as scalars representing the total sample size and class VC-dimension,

[2] In that case, technically, if there does not exists a k in A such that $g(k) = \inf_{k'} g(k')$ then we can always find an arbitrarily close approximating elements k_n, i.e., $\forall \epsilon > 0$ $\exists N(\epsilon)$ such that for $n > N(\epsilon)$ we have $|g(k_n) - \inf_{k'} g(k')| < \epsilon$.

respectively. Let us denote by $\mathcal{C}_k$ a class of classifiers having a VC-dimension of k and let c_k^* be the classifier which minimises the loss $L(c)$ over $\mathcal{C}_k$, i.e., $c_k^* = \text{argmin}_{c \in \mathcal{C}_k} L(c)$. The standard setting for SRM considers the overall class $\mathcal{C}$ of classifiers as an infinite union of finite VC-dimension classes, i.e., $\bigcup_{k=1}^{\infty} \mathcal{C}_k$, see for instance Vapnik [1982], Devroye et. al. [1996], Shawe-Taylor et. al. [1996], Lugosi & Nobel [1996], Ratsaby et. al. [1996]. The best performing classifier in $\mathcal{C}$ denoted as c^* is defined as $c^* = \text{argmin}_{1 \leq k \leq \infty} L(c_k^*)$. Similarly, denote by $\hat{c}_k$ the empirically-best classifier in $\mathcal{C}_k$, i.e., $\hat{c}_k = \text{argmin}_{c \in \mathcal{C}_k} L_m(c)$. Denoting by k^* the *minimal complexity* of a class which contains c^*, then depending on the problem and on the type of classifiers used, k^* may even be infinite as in the case when the Bayes classifier is not contained in $\mathcal{C}$. The complexity k^* may be thought of as the intrinsic complexity of the Bayes classifier.

The idea behind SRM is to minimise not the pure empirical loss $L_m(c_k)$ but a penalised version taking the form $L_m(c_k) + \epsilon(m,k)$ where $\epsilon(m,k)$ is some increasing function of k and is sometimes referred to as a *complexity penalty*. The classifier chosen by the criterion is then defined by

$$\hat{c}^* = \text{argmin}_{1 \leq k \leq \infty} \left(L_m(\hat{c}_k) + \epsilon(m,k) \right). \tag{1}$$

The term $\epsilon(m,k)$ is proportional to the worst case deviations between the true loss and the empirical loss uniformly over all functions in $\mathcal{C}_k$. More recently there has been interest in data-dependent penalty terms for structural risk minimisation which do not have an explicit complexity factor k but are related to the class $\mathcal{C}_k$ by being defined as a supremum of some empirical quantity over $\mathcal{C}_k$, for instance the maximum discrepancy criterion [Bartlett et. al., 2002] or the Rademacher complexity [Kultchinskii, 2002].

We take the penalty to be as in Vapnik [1982] (see also Devroye et. al. [1996]) $\epsilon(m,k) = \textit{const}\ \sqrt{\frac{k \ln m}{m}}$ where again *const* stands for an absolute constant which for our purpose is not important. This bound is central to the computations of the paper [3].

It can be shown [Devroye et. al., 1996] that for the two-pattern classification case the error rate of the SRM-chosen classifier $\hat{c}^*$ (which implicitly depends on the random sample of size m since it is obtained by minimising the sum in

[3] There is actually an improved bound due to Talagrand, cf. Anthony & Bartlett [1999] Section 4.6, but when adapted for almost sure statements it yields $O(\sqrt{\frac{k+\ln m}{m}})$ which for our work is insignificantly better then $O\left(\sqrt{\frac{k \ln m}{m}}\right)$

(1)), satisfies

$$L(\hat{c}^*) > L(c^*) + const\sqrt{\frac{k^* \ln m}{m}} \tag{2}$$

infinitely often with probability 0 where again c^* is the Bayes classifier which is assumed to be in $\mathcal{C}$ and k^* is its intrinsic complexity. The nice feature of SRM is that the selected classifier $\hat{c}^*$ automatically locks onto the minimal error rate as if the unknown k^* was *known* beforehand.

3 Multicategory classification

A classifier $c(x)$ may be represented as a vector of M boolean classifiers $b_i(x)$, where $b_i(x) = 1$ if x is a pattern drawn from class 'i' and $b_i(x) = 0$ otherwise. A union of such boolean classifiers forms a *well-defined classifier* $c(x)$ if for each $x \in \mathrm{I\!R}^d$, $b_i(x) = 1$ for exactly one i, i.e., $\bigcup_{i=1}^M \{x : b_i(x) = 1\} = \mathrm{I\!R}^d$ and $\{x : b_i(x) = 1\} \bigcap \{x : b_j(x) = 1\} = \emptyset$, for $1 \leq i \neq j \leq M$. We also refer to these boolean classifiers as the component classifiers $c_i(x)$, $1 \leq i \leq M$, of a vector classifier $c(x)$. The loss of a classifier c is just the average of the losses of the component classifiers, i.e., $L(c) = \sum_{i=1}^M p_i L(c_i)$ where for a boolean classifier c_i the loss is defined as $L(c_i) = \mathrm{E}_i 1_{\{c_i(x) \neq 1\}}$, and the empirical loss is $L_{i,m_i}(c_i) = \frac{1}{m_i} \sum_{j=1}^{m_i} 1_{\{c_i(x_j) \neq 1\}}$ which is based on a subsample $\{(x_j, i)\}_{j=1}^{m_i}$ drawn i.i.d. from pattern class "i".

The class $\mathcal{C}$ of classifiers is decomposed into a structure $S = S_1 \times S_2 \times \cdots \times S_M$, where S_i is a nested structure (cf. Vapnik [1982]) of classes $\mathcal{B}_{k_i}$, $i = 1, 2, \ldots$, of boolean classifiers $b_i(x)$, i.e., $S_1 = \mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k_1}, \ldots$, $S_2 = \mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k_2}, \ldots$ up to $S_M = \mathcal{B}_1, \mathcal{B}_2, \ldots, \mathcal{B}_{k_M}, \ldots$ where $k_i \in \mathbb{Z}_+$ denotes the VC-dimension of $\mathcal{B}_{k_i}$ and $\mathcal{B}_{k_i} \subseteq \mathcal{B}_{k_i+1}$, $1 \leq i \leq M$. For any fixed positive integer vector $k \in \mathbb{Z}_+^M$ consider the class of vector classifiers $\mathcal{C}_k = \mathcal{B}_{k_1} \times \mathcal{B}_{k_2} \times \cdots \times \mathcal{B}_{k_M}$. Define by $\mathcal{G}_k$ the subclass of $\mathcal{C}_k$ of classifiers c that are well-defined (in the sense mentioned above).

For vectors m and k in $\mathbb{Z}_+^M$, define $\epsilon(m, k) \equiv \sum_{i=1}^M p_i \epsilon(m_i, k_i)$ where as before $\epsilon(m_i, k_i) = const\ \sqrt{\frac{k_i \ln m_i}{m_i}}$. For any $0 < \delta < 1$, we denote by $\epsilon(m_i, k_i, \delta) = \sqrt{\frac{k_i \ln m_i + \ln \frac{1}{\delta}}{m_i}}$ and $\epsilon(m, k, \delta) = \sum_{i=1}^M p_i \epsilon(m_i, k_i, \delta)$. Lemma 1 below states an upper bound on the deviation between the empirical loss and the loss uniformly over all classifiers in a class $\mathcal{G}_k$ and is a direct application of Theorem 6.7 Vapnik [1982]. Before we state it, it is necessary to define what is meant by an increasing sequence of vectors m.

Definition 1 (Increasing sample-size sequence) *A sequence $m(n)$ of sample-*

size vectors is said to increase if: (a) at every n, there exists a j such that $m_j(n+1) > m_j(n)$ and $m_i(n+1) \geq m_i(n)$ for $1 \leq i \neq j \leq M$ and (b) there exists an increasing function $T(N)$ such that for all $N > 0$, $n > N$ implies every component $m_i(n) > T(N)$, $1 \leq i \leq M$.

Definition 1 implies that for all $1 \leq i \leq M$, $m_i(n) \to \infty$ as $n \to \infty$. We will henceforth use the notation $m \to \infty$ to denote such an ever-increasing sequence $m(n)$ with respect to an implicit discrete indexing variable n. The relevance of Definition 1 will become clearer later, in particular when considering Lemma 3.

Definition 2 (Sequence generating procedure) *A* sequence generating procedure *ϕ is one which generates increasing sequences $m(n)$ with a* fixed *function $T_\phi(N)$ as in Definition 1 and also satisfying the following: for all $N, N' \geq 1$ such that $T_\phi(N') = T_\phi(N)+1$ then $|N'-N| \leq const$, where const is dependent only on ϕ.*

The above definition simply states a lower bound requirement on the rate of increase of $T_\phi(N)$. We now state the uniform strong law of large numbers for the class of well-defined classifiers.

Lemma 1 (Uniform SLLN for multicategory classifier class) *For any $k \in \mathbb{Z}_+^M$ let $\mathcal{G}_k$ be a class of well-defined classifiers. Consider any sequence-generating procedure as in Definition 2 which generates $m(n)$, $n = 1, \ldots, \infty$. Let the empirical loss be defined based on examples $\{(x_j, y_j)\}_{j=1}^{\overline{m}(n)}$, each drawn i.i.d. according to an unknown underlying distribution over $\mathbb{R}^d \times \{1, \ldots, M\}$. Then for arbitrary $0 < \delta < 1$, $\sup_{c \in \mathcal{G}_k} \left| L_{m(n)}(c) - L(c) \right| \leq const\ \epsilon(m(n), k, \delta)$ with probability $1-\delta$ and the events $\sup_{c \in \mathcal{G}_k} \left| L_{m(n)}(c) - L(c) \right| > const\ \epsilon(m(n), k)$, $n = 1, 2, \ldots$, occur infinitely often with probability 0, where $m(n)$ is any sequence generated by the procedure.*

The outline of the proof is in Appendix A. We henceforth denote by c_k^* the optimal classifier in $\mathcal{G}_k$, i.e., $c_k^* = \mathrm{argmin}_{c \in \mathcal{G}_k} L(c)$ and $\hat{c}_k = \mathrm{argmin}_{c \in \mathcal{G}_k} L_m(c)$ is the empirical minimiser over the class $\mathcal{G}_k$.

In Section 2 we mentioned that the intrinsic unknown complexity k^* of the Bayes classifier is automatically learned by minimising the penalised empirical loss over the complexity variable k. If an upper bound of the form of (2) but based on a vector m could be derived for the multicategory case then a second minimisation step, this time over the sample-size vector m, will improve the SRM error convergence rate. The main result of this paper (Theorem 1) shows that through a stochastic gradient descent such minimisation improves the standard SRM bound from $\epsilon(m, k^*)$ to $\epsilon(m^*, k^*)$ where m^* minimises $\epsilon(m, k^*)$ over all possible vectors m whose magnitude $\|m\|$ equals the given total sample size $\overline{m}$. The technical challenge is to obtain this without assuming the knowledge of k^*. Our approach is to estimate k^* and minimise an estimated criterion. Due to lack of space, we only provide sketch of proofs for the stated lemmas and theorem. The full proofs will appear in the full paper [Ratsaby, 2003].

Concerning the convergence mode of random variables, upper bounds are based on the uniform strong law of large numbers, see Appendix A. Such bounds originated in the work of Vapnik [1982], for instance his Theorem 6.7. Throughout the current paper, *almost sure* statements are made by a standard application of the Borel-Cantelli lemma. For instance, taking m to be a scalar, the statement $\sup_{b \in B} |L(b) - L_m(b)| \le const\sqrt{\frac{r \log m + \log \frac{1}{\delta}}{m}}$ with probability at least $1 - \delta$ for any $\delta > 0$ is alternatively stated as follows by letting $\delta_m = \frac{1}{m^2}$: For the sequence of random variables $L_m(b)$, uniformly over all $b \in B$, we have $L(b) > L_m(b) + const\sqrt{\frac{r \log m + \log \frac{1}{\delta_m}}{m}}$ occur infinitely often with probability 0. Concerning our, perhaps loose, use of the word *optimal*, whenever not explicitly stated, optimality of a classifier or of a procedure or algorithm is only with respect to minimisation of the criterion, namely, the upper bound on the loss.

4 Standard SRM Loss Bounds

We will henceforth make the following assumption.

Assumption 1 *The Bayes loss $L^* = 0$ and there exists a classifier c_k in the structure S with $L(c_k) = L^*$ such that $k_i < \infty$, $1 \le i \le M$. The* a priori *pattern class probabilities p_i, $1 \le i \le M$, are known to the learner.*

Assumption 1 essentially amounts to the Probably Approximately Correct (PAC) framework, Valiant [1984], Devroye et. al. [1996] Section 12.7, but with a more relaxed constraint on the complexity of the hypothesis class $\mathcal{C}$ since it is permitted to have an infinite VC-dimension. Also, in practice the *a priori* pattern class probabilities can be estimated easily. In assuming that the learner knows the p_i, $1 \le i \le M$, one approach would have the learner allocate sub-sample sizes according to $m_i = p_i \overline{m}$ followed by doing structural risk minimisation. However this does not necessarily minimise the upper bound on the loss of the SRM-selected classifier and hence is inferior in this respect to Principle 1 which is stated later. We note that if the classifier class was *fixed* and the intrinsic complexity k^* of the Bayes classifier was known in advance then because of Assumption 1 one would resort to a bound of the form $O\left(k \log m/m\right)$ and not the weaker bound that has a square root, see ch. 4.5 in Anthony & Bartlett [1999]. However, as mentioned before, not knowing k^* and hence using structural risk minimisation as opposed to empirical risk minimisation over a fixed class, leads to using the weaker bound for the complexity-penalty.

We next provide some additional definitions needed for the remainder of the paper. Consider the set $F^* = \{\mathrm{argmin}_{k \in \mathbb{Z}_+^M} L(c_k^*)\} = \{k : L(c_k^*) = L^* = 0\}$ which may contain more than one vector k. Following Assumption 1 we may define the *Bayes classifier* c^* as the particular classifier $c_{k^*}^*$ whose complexity is minimal, i.e., $k^* = \mathrm{argmin}_{\{k \in F^*\}}\{\|k\|_\infty\}$ where $\|k\|_\infty = \max_{1 \leq i \leq M} |k_i|$. Note again that there may be more than one such k^*. The significance of specifying the Bayes classifier up to its complexity rather than just saying it is any classifier having a loss L^* will become apparent later in the paper.

For an empirical minimiser classifier $\hat{c}_k$ define by the *penalised empirical loss* (cf. Devroye et. al. [1996]) $\tilde{L}_m(\hat{c}_k) = L_m(\hat{c}_k) + \epsilon(m, k)$. Consider the set $\hat{F} = \{\mathrm{argmin}_{k \in \mathbb{Z}_+^M} \tilde{L}(\hat{c}_k)\}$ which may contain more than one vector k. In standard structural risk minimisation [Vapnik, 1982] the selected classifier is *any* one whose complexity index $k \in \hat{F}$. This will be modified later when we introduce an algorithm which relies on the convergence of the complexity $\hat{k}$ to some finite limiting complexity value with increasing[4] m. The selected classifier is therefore defined to be one whose complexity satisfies $\hat{k} = \mathrm{argmin}_{k \in \hat{F}} \|k\|_\infty$. This *minimal-complexity SRM-selected classifier* will be denoted as $\hat{c}_{\hat{k}}$ or simply as $\hat{c}^*$. We will sometimes write $\hat{k}_n$ and $\hat{c}_n$ for the complexity and for the SRM-selected classifier, respectively, in order to explicitly show the dependence on discrete time n.

The next lemma states that the complexity $\hat{k}$ converges to some (not necessarily unique) k^* corresponding to the Bayes classifier c^*.

Lemma 2 *Based on $\overline{m}$ examples $\{(x_j, y_j)\}_{j=1}^{\overline{m}}$ each drawn i.i.d. according to an unknown underlying distribution over $\mathbb{R}^d \times \{1, \ldots, M\}$, let $\hat{c}^*$ be the chosen classifier of complexity $\hat{k}$. Consider a sequence of samples $\zeta^{m(n)}$ with increasing sample-size vectors $m(n)$ obtained by a sequence-generating procedure as in Definition 2. Then (a) the corresponding complexity sequence $\hat{k}_n$ converges a.s. to k^* which from Assumption 1 has finite components. (b) For any sample $\zeta^{m(n)}$ in the sequence, the loss of the corresponding classifier $\hat{c}_n^*$ satisfies $L(\hat{c}_n^*) >$ const $\epsilon(m(n), k^*)$ infinitely often with probability* 0.

The outline of the proof is in Appendix B. For the more general case of $L^* > 0$ (but two-category classifiers) the upper bound becomes $L^* + \mathit{const}\ \epsilon(m, k^*)$, cf. Devroye et. al. [1996]. It is an open question whether in this case it is possible to guarantee convergence of $\hat{k}_n$ or some variation of it to a finite limiting value.

[4] We will henceforth adopt the convention that a vector sequence $\hat{k}_n \to k^*$, a.s., means that every component of $\hat{k}_n$ converges to the corresponding component of k^*, a.s., as $m \to \infty$.

The previous lemma bounds the loss of the SRM-selected classifier $\hat{c}^*$. As suggested earlier, we wish to extend the SRM approach to do an additional minimisation step by minimising the loss of $\hat{c}^*$ with respect to the sample size vector m. In this respect, the subsample proportions may be tuned to the intrinsic Bayes complexity k^* thereby yield an improved error rate for $\hat{c}^*$. This is stated next:

Principle 1 *Choose m to minimise the criterion $\epsilon(m, k^*)$ with respect to all m such that $\sum_{i=1}^{M} m_i = \overline{m}$, the latter being the a priori total sample size allocated for learning.*

In general there may be other proposed criteria just as there are many criteria for model selection based on minimisation of different upper bounds. Note that if k^* was known then an optimal sample size $m^* = [m_1^*, \ldots, m_M^*]$ could be computed which yields a classifier $\hat{c}^*$ with the best (lowest) deviation *const* $\epsilon(m^*, k^*)$ away from Bayes loss. The difficulty is that $k^* = [k_1^*, \ldots, k_M^*]$ is usually unknown since it depends on the underlying unknown probability densities $f_i(x)$, $1 \leq i \leq M$. To overcome this we will minimise an estimate of $\epsilon(\cdot, k^*)$ rather than the criterion $\epsilon(\cdot, k^*)$ itself.

5 The Extended SRM algorithm

In this section we extend the SRM learning algorithm to include a stochastic gradient descent step. The idea is to interleave the standard minimisation step of SRM with a new step which asymptotically minimises the penalised empirical loss with respect to the sample size. As before, $m(n)$ denotes a sequence of sample-size vectors indexed by an integer $n \geq 0$ representing discrete time. When referring to a particular i^{th} component of the vector $m(n)$ we write $m_i(n)$.

The algorithm initially starts with uniform sample size proportions $m_1 = m_2 = \cdots = m_M = const > 0$, then at each time $n \geq 1$ it selects the classifier $\hat{c}_n^*$ defined as

$$\hat{c}_n^* = \operatorname{argmin}_{\hat{c}_{n,k}: k \in \hat{F}_n} \|k\|_\infty \qquad \textbf{Standard Minimization Step} \qquad (3)$$

where $\hat{F}_n = \{k : \tilde{L}_n(\hat{c}_{n,k}) = \min_{r \in \mathbb{Z}_+^M} \tilde{L}_n(\hat{c}_{n,r})\}$ and for any $\hat{c}_{n,k}$ which minimises $L_{m(n)}(c)$ over all $c \in \mathcal{G}_k$ we define the penalised empirical loss as $\tilde{L}_n(\hat{c}_{n,k}) = L_{m(n)}(\hat{c}_{n,k}) + \epsilon(m(n), k)$ where $L_{m(n)}$ stands for the empirical loss based on the sample-size vector $m(n)$ at time n.

The second minimisation step is done via a query rule which selects the particular pattern class from which to draw examples as one which minimises the stochastic criterion $\epsilon(\cdot, \hat{k}_n)$ with respect to the sample size vector $m(n)$. The complexity $\hat{k}_n$ of $\hat{c}_n^*$ will be shown later to converge to k^* hence $\epsilon(\cdot, \hat{k}_n)$ serves as a consistent estimator of the criterion $\epsilon(\cdot, k^*)$. We choose an adaptation step which changes one component of m at a time, namely, it increases the component $m_{j_{max}(n)}$ which corresponds to the direction of maximum descent of the criterion $\epsilon(\cdot, \hat{k}_n)$ at time n. This may be written as

$$m(n+1) = m(n) + \Delta\, e_{j_{max}} \qquad \textbf{New Minimization Step} \tag{4}$$

where the positive integer Δ denotes some fixed minimisation step-size and for any integer $i \in \{1, 2, \ldots, M\}$, e_i denotes an M-dimensional elementary vector with 1 in the i^{th} component and 0 elsewhere. Thus at time n the new minimisation step produces a new value $m(n+1)$ which is used for drawing additional examples according to specific sample sizes $m_i(n+1)$, $1 \leq i \leq M$.

Learning Algorithm XSRM (Extended SRM)
Let: $m_i(0) = const > 0$, $1 \leq i \leq M$.
Given: (a) M uniform-size samples $\{\zeta^{m_i(0)}\}_{i=1}^M$, where $\zeta^{m_i(0)} = \{(x_j, \text{‘i’})\}_{j=1}^{m_i(0)}$, and x_j are drawn i.i.d. according to underlying class-conditional probability densities $f_i(x)$. (b) A sequence of classes $\mathcal{G}_k$, $k \in \mathbb{Z}_+^M$, of well-defined classifiers. (c) A constant minimisation step-size $\Delta > 0$. (d) Known *a priori* probabilities p_j, $1 \leq j \leq M$ (for defining L_m).
Initialisation: (Time $n = 0$) Based on $\zeta^{m_i(0)}$, $1 \leq i \leq M$, determine a set of candidate classifiers $\hat{c}_{0,k}$ minimising the empirical loss $L_{m(0)}$ over $\mathcal{G}_k$, $k \in Z_+^M$, respectively. Determine $\hat{c}_0^*$ according to (3) and denote its complexity vector by $\hat{k}_0$.
Output: $\hat{c}_0^*$.
Call Procedure NM: $m(1) := NM(0)$.
Let $n = 1$**.**
While (still more available examples) **Do:**

1. Based on the sample $\zeta^{m(n)}$, determine the empirical minimisers $\hat{c}_{n,k}$ for each class $\mathcal{G}_k$. Determine $\hat{c}_n^*$ according to (3) and denote its complexity vector by $\hat{k}_n$.
2. **Output:** $\hat{c}_n^*$.
3. **Call Procedure NM:** $m(n+1) := NM(n)$.
4. $n := n + 1$.

End Do

□

Procedure New Minimisation (NM)
Input: Time n.

- $j_{max}(n) := \text{argmax}_{1 \leq j \leq M}\ p_j \frac{\epsilon(m_j(n), \hat{k}_{n,j})}{m_j(n)}$, where if more than one argmax then choose any one.

- **Obtain:** Δ new i.i.d. examples from class $j_{max}(n)$. Denote them by ζ_n.
- **Update Sample:** $\zeta^{m_{j_{max}(n)}(n+1)} := \zeta^{m_{j_{max}(n)}(n)} \bigcup \zeta_n$, while $\zeta^{m_i(n+1)} := \zeta^{m_i(n)}$, for $1 \leq i \neq j_{max}(n) \leq M$.
- **Return Value:** $m(n) + \Delta\, e_{j_{max}(n)}$.

□

The algorithm alternates between the standard minimisation step (3) and the new minimisation step (4) repetitively until exhausting the total sample size $\overline{m}$ which for most generality is assumed to be unknown *a priori*.

While for any *fixed* $i \in \{1, 2, \ldots, M\}$ the examples $\{(x_j, i)\}_{j=1}^{m_i(n)}$ accumulated up until time n are all i.i.d. random variables, the total sample $\{(x_j, y_j)\}_{j=1}^{\overline{m}(n)}$ consists of *dependent* random variables since based on the new minimisation the choice of the particular class-conditional probability distribution used to draw examples at each time instant l depends on the sample accumulated up until time $l-1$. It turns out that this dependency does not alter the results of Lemma 2. This follows from the proof of Lemma 2 and from the bound of Lemma 1 which holds even if the sample is i.i.d. *only* when conditioned on a pattern class since it is the weighted average of the individual bounds corresponding to each of the pattern classes. Therefore together with the next lemma this implies that Lemma 2 applies to Algorithm XSRM.

Lemma 3 *Algorithm XSRM is a sequence-generating procedure.*

The outline of the proof is deferred to Appendix C. Next, we state the main theorem of the paper.

Theorem 1 *Assume that the Bayes complexity k^* is an* unknown *M-dimensional vector of finite positive integers. Let the step size $\Delta = 1$ in Algorithm XSRM resulting in a total sample size which increases with discrete time as $\overline{m}(n) = n$. Then the random sequence of classifiers $\hat{c}_n^*$ produced by Algorithm XSRM is such that the events $L(\hat{c}_n^*) > const\ \epsilon(m(n), k^*)$ or $\|m(n) - m^*(n)\|_{l_1^M} > 1$ occur infinitely often with probability 0 where $m^*(n)$ is the solution to the constrained minimisation of $\epsilon(m, k^*)$ over all m of magnitude $\|m\| = \overline{m}(n)$.*

Remark 1 *In the limit of large n the bound const $\epsilon(m(n), k^*)$ is almost minimum (the minimum being at $m^*(n)$) with respect to all vectors $m \in \mathbb{Z}_+^M$ of size $\overline{m}(n)$. Note that this rate is achieved by Algorithm XSRM without the knowledge of the intrinsic complexity k^* of the Bayes classifier. Compare this for instance to uniform querying where at each time n one queries for subsamples of the same size $\frac{\Delta}{M}$ from* every *pattern class. This leads to a different (deterministic) sequence $m(n) = \frac{\Delta}{M}[1, 1, \ldots, 1]n \equiv \underline{\Delta}\, n$ and in turn to a sequence of classifiers $\hat{c}_n$ whose loss $L(\hat{c}_n) \leq const\ \epsilon(\underline{\Delta}\, n, k^*)$, as $n \to \infty$, where here the upper bound is not even asymptotically minimal. A similar argument holds* *if the proportions are based on the* a priori *pattern class probabilities since in general letting $m_i = p_i \overline{m}$ does not necessarily minimise the upper bound. In Ratsaby [1998], empirical results show the inferiority of uniform sampling compared to an online approach based on Algorithm XSRM.*

6 Proving Theorem 1

The proof of Theorem 1 is based on Lemma 2 and on two additional lemmas, Lemma 4 and Lemma 5, which deal with the the convergent property of the new minimisation step of Algorithm XSRM. The proof is outlined in Appendix D. Our approach is to show that the adaptation step used in the new minimisation step follows from the minimisation of the deterministic criterion $\epsilon(m, k^*)$ with a known k^*. Letting t, as well as n, denote discrete time $t = 1, 2, \ldots$, we adopt the notation $m(t)$ for a *deterministic* sample size sequence governed by the deterministic criterion $\epsilon(m, k^*)$ where k^* is taken to be known. We write $m(n)$ to denote the *stochastic* sequence governed by the random criterion $\epsilon(m, \hat{k}_n)$. Thus t or n distinguish between a deterministic or stochastic sample sequence, $m(t)$ or $m(n)$, respectively. We start with the following definition.

Definition 3 (Optimal trajectory) *Let $\overline{m}(t)$ be any positive integer-valued function of t which denotes the total sample size at time t. The optimal trajectory is a set of vectors $m^*(t) \in \mathbb{Z}_+^M$ indexed by $t \in \mathbb{Z}_+$, defined as $m^*(t) = argmin_{\{m \in \mathbb{Z}_+^M : \|m\| = \overline{m}(t)\}} \epsilon(m, k^*)$.*

First let us solve the following constrained minimisation problem. Fix a total sample size $\overline{m}$ and minimise the error $\epsilon(m, k^*)$ under the constraint that $\sum_{i=1}^M m_i = \overline{m}$. This amounts to minimising $\epsilon(m, k^*) + \lambda(\sum_{i=1}^M m_i - \overline{m})$ over m and λ. Denote the gradient by $g(m, k^*) = \nabla \epsilon(m, k^*)$. Then the above is equivalent to solving $g(m, k^*) + \lambda[1, 1, \ldots, 1] = 0$ for m and λ. The vector valued function $g(m, k^*)$ may be approximated by $g(m, k^*) \simeq [-\frac{p_1 \epsilon(m_1, k_1^*)}{2m_1}, -\frac{p_2 \epsilon(m_2, k_2^*)}{2m_2}, \ldots, -\frac{p_M \epsilon(m_M, k_M^*)}{2m_M}]$ where we used the approximation $1 - \frac{1}{\log m_i} \simeq 1$ for $1 \leq i \leq M$. We then obtain the set of equations $2\lambda^* m_i^* = p_i \epsilon(m_i^*, k_i^*)$, $1 \leq i \leq M$, and $\lambda^* = \frac{\epsilon(m^*, k^*)}{2\overline{m}}$. We are interested not in obtaining a solution for a fixed $\overline{m}$ but obtaining, using local gradient information, a sequence of solutions for the sequence of minimization problems corresponding to an increasing sequence of total sample-size values $\overline{m}(t)$.

Applying the New Minimization procedure of Algorithm XSRM to the deterministic criterion $\epsilon(m, k^*)$ we have an adaptation rule which modifies the sample size vector $m(t)$ at time t in the direction of steepest descent of $\epsilon(m, k^*)$. This yields: $j^*(t) = \text{argmax}_{1 \leq j \leq M} \frac{p_j \epsilon(m_j(t), k_j^*)}{m_j(t)}$ which means we let

$m_{j^*(t)}(t+1) = m_{j^*(t)}(t) + \Delta$ while the remaining components of $m(t)$ remain unchanged, i.e., $m_j(t+1) = m_j(t), \forall j \neq j^*(t)$. The next lemma states that this rule achieves the desired result, namely, the deterministic sequence $m(t)$ converges to the optimal trajectory $m^*(t)$.

Lemma 4 *For any initial point $m(0) \in \mathbb{R}^M$, satisfying $m_i(0) \geq 3$, for a fixed positive Δ, there exists some finite integer $0 < N' < \infty$ such that for all discrete time $t > N'$ the trajectory $m(t)$ corresponding to a repeated application of the adaptation rule $m_{j^*(t)}(t+1) = m_{j^*(t)}(t) + \Delta$ is no farther than Δ (in the l_1^M-norm) from the optimal trajectory $m^*(t)$.*

Outline of Proof: Recall that $\epsilon(m, k^*) = \sum_{i=1}^M p_i \epsilon(m_i, k_i^*)$ where $\epsilon(m_i, k_i) = \sqrt{\frac{k_i \ln m_i}{m_i}}$, $1 \leq i \leq M$. The derivative $\frac{\partial \epsilon(m,k^*)}{\partial m_i} \simeq -p_i \frac{\epsilon(m_i,k_i^*)}{2m_i}$. Denote by $x_i = p_i \frac{\epsilon(m_i,k_i^*)}{2m_i}$, and note that $\frac{dx_i}{dm_i} \simeq -\frac{3}{2}\frac{x_i}{m_i}$, $1 \leq i \leq M$. There is a one-to-one correspondence between the vector x and m thus we may refer to the optimal trajectory also in x-space. Consider the set $T = \{x = c[1, 1, \ldots, 1] \in \mathbb{R}_+^M : c \in \mathbb{R}_+\}$ and refer to T' as the corresponding set in m-space. Define the Lyapunov function $V(x(t)) = V(t) = \frac{x_{max}(t) - x_{min}(t)}{x_{min}(t)}$ where for any vector $x \in \mathbb{R}_+^M$, $x_{max} = \max_{1 \leq i \leq M} x_i$, and $x_{min} = \min_{1 \leq i \leq M} x_i$, and write m_{max}, m_{min} for the elements of m with the same index as x_{max}, x_{min}, respectively. Denote by $\dot{V}$ the derivative of V with respect to t. Using standard analysis it can be shown that if $x \notin T$ then $V(x) > 0$ and $\dot{V}(x) < 0$ while if $x \in T$ then $V(x) = 0$ and $\dot{V}(x) = 0$. This means that as long as $m(t)$ is not on the optimal trajectory then $V(t)$ decreases. To show that the trajectory is an attractor $V(t)$ is shown to decrease fast enough to zero using the fact that $V(t) \leq const \left(\frac{1}{t}\right)^{\frac{3}{2}}$. Hence as $t \to \infty$, the distance between $m(t)$ and the set T' $\mathrm{dist}(m(t), T') \to 0$ where $\mathrm{dist}(x, T) = \inf_{y \in T} \|x - y\|_{l_1^M}$ and l_1^M denotes the Euclidean vector norm. It is then easy to show that for all large t, $m(t)$ is farther from $m^*(t)$ by no more than Δ. □

We now show that the same adaptation rule may also be used in the setting where k^* is unknown. The next lemma states that even when k^* is unknown, it is possible, by using Algorithm XSRM, to generate a stochastic sequence which asymptotically converges to the optimal $m^*(n)$ trajectory (again, the use of n instead of t just means we have a random sequence $m(n)$ and not a deterministic sequence $m(t)$ as was investigated above).

Lemma 5 *Fix any $\Delta \geq 1$ as a step size used by Algorithm XSRM. Given a sample size vector sequence $m(n)$, $n \to \infty$, generated by Algorithm XSRM, assume that $\hat{k}_n \to k^*$ almost surely. Let $m^*(n)$ be the optimal trajectory as in Definition 3. Then the events $\|m(n) - m^*(n)\|_{l_1^M} > \Delta$ occur infinitely often with probability* 0.

Outline of Proof: From Lemma 3 $m(n)$ generated by Algorithm XSRM is an increasing sample-size sequence. Therefore by Lemma 2 we have $\hat{k}_n \to k^*$, a.s., as $n \to \infty$. This means that $P(\exists n > N, |\hat{k}_n - k^*| > \epsilon) = \delta_N(\epsilon)$ where $\delta_N(\epsilon) \to 0$ as $N \to \infty$. It follows that for all $\delta > 0$, there is a finite $N(\delta, \epsilon) \in \mathbb{Z}_+$ such that with probability $1 - \delta$ for all $n \geq N(\epsilon, \delta)$, $\hat{k}_n = k^*$. It follows that with the same probability for all $n \geq N$, the criterion $\epsilon(m, \hat{k}_n) = \epsilon(m, k^*)$, uniformly over all $m \in \mathbb{Z}_+^M$, and hence the trajectory $m(n)$ taken by Algorithm XSRM, governed by the criterion $\epsilon(\cdot, \hat{k}_n)$, equals the trajectory $m(t)$, $t \in \mathbb{Z}_+$, taken by minimising the deterministic criterion $\epsilon(\cdot, k^*)$. Moreover, this probability of $1 - \delta$ goes to 1 as $N \to \infty$ by the a.s. convergence of $\hat{k}_n$ to k^*. By Lemma 4, there exists a $N' < \infty$ such that for all discrete time $t > N'$, $\|m(t) - m^*(t)\|_{l_1^M} \leq \Delta$. Let $N'' = \max\{N, N'\}$ then $P\left(\exists n > N'', \hat{k}_n \neq k^* \text{ or } \left\|m(t)_{|t=n} - m^*(t)_{|t=n}\right\|_{l_1^M} > \Delta\right) = \delta_{N''}$ where $\delta_{N''} \to 0$ as $N'' \to \infty$. The latter means that the event $\hat{k}_n \neq k^*$ or $\|m(n) - m^*(n)\|_{l_1^M} > \Delta$ occurs infinitely often with probability 0. The statement of the lemma then follows. □

Appendix

Due to space limitation only the outline of the proofs is included. Complete proofs are available in the full paper on-line[5].

A Outline of Proof of Lemma 1

For a class of boolean classifiers $\mathcal{B}_r$ of VC-dimension r it is known (cf. Devroye et. al. [1996] ch. 6, Vapnik [1982] Theorem 6.7) that a bound on the deviation between the loss and the empirical loss uniformly over all classifiers $b \in \mathcal{B}_r$ is $\sup_{b\in\mathcal{B}_r} |L(b) - L_m(b)| \le const \sqrt{\frac{r \ln m + \ln\left(\frac{1}{\delta}\right)}{m}}$ with probability $1 - \delta$ where m denotes the size of the random sample used for calculating empirical loss $L_m(b)$. Choosing for instance $\delta_m = \frac{1}{m^2}$ implies that the bound $const\sqrt{\frac{r \ln m}{m}}$ (with a different constant) does not hold infinitely often with probability 0. We will refer to this as the uniform strong law of large numbers result and we note that this was defined earlier as $\epsilon(m, r)$.

This result is used together with an application of the union bound which reduces the probability $\mathbf{P}\left(\sup_{c\in\mathcal{C}_k} |L(c) - L_m(c)| > \epsilon(m, k, \delta')\right)$ into $\sum_{i=1}^{M} \mathbf{P}(\exists c \in \mathcal{C}_{k_i} : |L(c) - L_{i,m_i}(c)| > \epsilon(m_i, k_i, \delta'))$ which is bounded from above by $M\delta'$. The first part of the lemma then follows since the class of well defined classifiers $\mathcal{G}_k$ is contained in the class $\mathcal{C}_k$. For the second part of the lemma, by the premise consider any fixed complexity vector k and any sequence-generating procedure ϕ according to Definition 2. Define the following set of sample size vector sequences: $A_N \equiv \{m(n) : n > N, m(n) \text{ is generated by } \phi\}$. As the space is discrete, note that for any finite N, the set A_N contains all possible paths except a finite number of length-N paths. The proof proceeds by showing that the events $E_n \equiv \{\sup_{c\in\mathcal{G}_k} \left|L(c) - L_{m(n)}(c)\right| > \epsilon(m(n), k, \delta) : m(n) \text{ generated by } \phi\}$ occur infinitely often with probability 0. To show this, we first choose for δ to be $\delta^*_m = \frac{1}{\max_{1\le j\le M} m_j^2}$, and then reduce the $\mathbf{P}(\exists m(n) \in A_N : \sup_{c\in\mathcal{G}_k} \left|L(c) - L_{m(n)}(c)\right| > \epsilon(m(n), k, \delta^*_{m(n)}))$ to $\sum_{j=1}^{M} \sum_{m_j > T_\phi(N)} \frac{1}{m_j^2}$. Then use the fact that $m(n) \in A_N$ implies there exists a point m such that $\min_{1\le j\le M} m_j > T_\phi(N)$ where $T_\phi(N)$ is increasing with N hence the set $\{m_j : m_j > T_\phi(N)\}$ is strictly increasing, $1 \le j \le M$, which implies that the above double sum strictly decreases with increasing N. It then follows that $\lim_{N\to\infty} \mathbf{P}(\exists m(n) \in A_N : \sup_{c\in\mathcal{G}_k} \left|L(c) - L_{m(n)}(c)\right| > \epsilon(m(n), k)) = 0$ which implies the events E_n occur *i.o.* with probability 0. $\square$

[5] `http://www.cs.ucl.ac.uk/staff/J.Ratsaby/Publications/PDF/m-class.pdf`

B Outline of the Proof of Lemma 2

First we sketch the proof of the convergence of $\hat{k} \to k^*$, where k^* is some vector of minimal norm over all vectors k for which $L(c_k^*) = 0$. We henceforth denote for a vector $k \in \mathbb{Z}_+^M$, by $\|k\|_\infty = \max_{1 \le i \le M} |k_i|$. All convergence statements are made with respect to the increasing sequence $m(n)$. The indexing variable n is sometimes left hidden for simpler notation.

The set $\hat{F}$ defined in Section 4 may be rewritten as $\hat{F} = \{k : \tilde{L}(\hat{c}_k) = \tilde{L}(\hat{c}^*)\}$. The cardinality of $\hat{F}$ is finite since for all k having at least one component k_i larger than some constant implies $\tilde{L}(\hat{c}_k) > \tilde{L}(\hat{c}^*)$ because $\epsilon(m, k)$ will be larger than $\tilde{L}(\hat{c}^*)$ which implies that the set of k for which $\tilde{L}(\hat{c}_k) \le \tilde{L}(\hat{c}^*)$ is finite. Now for any $\alpha > 0$, define $\hat{F}_\alpha = \{k : \tilde{L}(\hat{c}_k) \le \tilde{L}(\hat{c}^*) + \alpha\}$. Recall that F^* was defined in Section 4 as $F^* = \{k : L(c_k^*) = L^* = 0\}$ and define $F_\alpha^* = \{k : L(c_k^*) \le L^* + \alpha\}$, where the Bayes loss is $L^* = 0$. Recall that the chosen classifier $\hat{c}^*$ has a complexity $\hat{k} = \text{argmin}_{k \in \hat{F}} \|k\|_\infty$. By Assumption 1, there exists a $k^* = \text{argmin}_{k \in F^*} \|k\|_\infty$ all of whose components are finite. The proof proceeds by first showing that $\hat{F} \not\subseteq F^*_{\epsilon(m,k^*)}$, i.o. with probability 0. Then proving that $k^* \in \hat{F}$ and that for all m large enough, $k^* = \text{argmin}_{k \in F^*_{\epsilon(m,k^*)}} \|k\|_\infty$. It then follows that $\|\hat{k}\|_\infty \ne \|k^*\|_\infty$ *i.o.* with probability zero but where $\hat{k}$ does not necessarily equal k^* and that $\hat{k} \to k^*$, (componentwise) *a.s.*, $m \to \infty$ (or equivalently, with $n \to \infty$ as the sequence $m(n)$ is increasing) where $k^* = \text{argmin}_{k \in F^*} \|k\|_\infty$ is not necessarily unique but all of whose components are finite. This proves the first part of the lemma. The proof of the second part of the Lemma follows similarly as the proof of Lemma 1. Start with $\mathbf{P}\left(\exists m(n) \in A_N : L(\hat{c}_n^*) > \epsilon(m(n), k^*)\ \right)$ which after some manipulation is shown to be bounded from above by the sum $\sum_{j=1}^M \sum_{k_j=1}^\infty \mathbf{P}\left(\exists m_j > T_\phi(N) : L(\hat{c}_{k_j}) > L_{j,m_j}(\hat{c}_{k_j}) + \epsilon(m_j, k_j)\right)$. Then make use of the uniform strong law result (see first paragraph of Appendix A) and choose a *const* such that $\epsilon(m_j, k_j) = const\sqrt{\frac{k_j \ln m_j}{m_j}} \ge \sqrt{3}\sqrt{\frac{k_j \ln(e m_j)}{m_j}}$. Using the upper bound on the growth function cf. Vapnik [1982] Section 6.9, Devroye et. al. [1996] Theorem 13.3, we have for some absolute constant $\kappa > 0$, $\mathbf{P}(L(\hat{c}_{k_j}) > L_{j,m_j}(\hat{c}_{k_j}) + \epsilon(m_j, k_j)) \le \kappa m_j^{k_j} e^{-m_j \epsilon^2(m_j, k_j)}$ which is bounded from above by $\kappa \frac{1}{m_j^2} e^{-3k_j}$ for $k_j \ge 1$. The bound on the double sum then becomes $2\kappa \sum_{j=1}^M \sum_{m_j > T_\phi(N)} \frac{1}{m_j^2}$ which is strictly decreasing with N as in the proof of Lemma 1. It follows that the events $\{L(\hat{c}_n^*) > \epsilon(m(n), k^*)\}$ occur infinitely often with probability 0. □

C Outline of the Proof of Lemma 3

Note that for this proof we cannot use Lemma 1 or parts of Lemma 2 since they are conditioned on having a sequence-generating procedure. Our approach here relies on the characteristics of the SRM-selected complexity $\hat{k}_n$ which is shown to be bounded uniformly over n based on Assumption 1. It follows that by the stochastic adaptation step of Algorithm XSRM the generated sample size sequence $m(n)$ is not only increasing but with a minimum rate of increase as in Definition 2. This establishes that Algorithm XSRM is a sequence-generating procedure. The proof starts by showing that for an increasing sequence $m(n)$, as in Definition 1, for all n there is some constant $0 < \rho < \infty$ such that $\|\hat{k}_n\|_\infty < \rho$. It then follows that for all n, $\hat{k}_n$ is bounded by a finite constant independent of n. So for a sequence generated by the new minimisation procedure in Algorithm XSRM, $p_j \frac{\epsilon(m_j(n),\hat{k}_{n,j})}{m_j(n)}$ are bounded by $p_j \frac{\epsilon(m_j(n),\tilde{k}_j)}{m_j(n)}$, for some finite $\tilde{k}_j$, $1 \leq j \leq M$, respectively. It can be shown by simple analysis of the function $\epsilon(m,k)$ that for a fixed k the ratio of $\frac{\partial^2 \epsilon(m_j,k_j)}{\partial m_j^2} / \frac{\partial^2 \epsilon(m_i,k_i)}{\partial m_i^2}$ converges to a constant dependent on k_i and k_j with increasing m_i, m_j. Hence the adaptation step which always increases one of the sub-samples yields increments of Δm_i and Δm_j which are no farther apart than a constant multiple of each other for all n, for any pair $1 \leq i,j \leq M$. Hence for a sequence $m(n)$ generated by Algorithm XSRM the following is satisfied: it is increasing in the sense of Definition 1, namely, for all $N > 0$ there exists a $T_\phi(N)$ such that for all $n > N$ every component $m_j(n) > T_\phi(N)$, $1 \leq j \leq M$. Furthermore, its rate of increase is bounded from below, namely, there exists a *const* > 0 such that for all $N, N' > 0$ satisfying $T_\phi(N') = T_\phi(N) + 1$, then $|N' - N| \leq$ *const*. It follows that Algorithm XSRM is a sequence-generating procedure according to Definition 2. $\square$

D Outline of Proof of Theorem 1

The classifier $\hat{c}_n^*$ is chosen according to (3) based on a sample of size vector $m(n)$ generated by Algorithm XSRM which is a sequence-generating procedure (by Lemma 3). From Lemma 2, $L(\hat{c}_n^*) >$ *const* $\epsilon(m(n), k^*)$, *i.o.* with probability 0 and since $\Delta = 1$ then from Lemma 5 it follows that $\|m(n) - m^*(n)\|_{l_1^M} > 1$ *i.o.* with probability 0 where $m^*(n) = \text{argmin}_{m:\|m\|=\overline{m}(n)} \epsilon(m, k^*)$. $\square$

References

Anthony M., Bartlett P. L., (1999), “Neural Network Learning:Theoretical Foundations”, Cambridge University Press, UK.

Bartlett P. L., Boucheron S., Lugosi G., (2002) Model Selection and Error Estimation, *Machine Learning*, Vol. 48(1-3), p. 85-113.

Devroye L., Gyorfi L. Lugosi G. (1996). “A Probabilistic Theory of Pattern Recognition”, Springer Verlag.

Kultchinskii V., (2002), Rademacher Penalties and Structural Risk Minimization, submitted to *IEEE Trans. on Info. Theory.*

Lugosi G., Nobel A., (1996), Adaptive Model Selection Using Empirical Complexities. Preprint, Dept. of Mathematics and Computer Sciences, Technical University of Budapest, Hungary.

Ratsaby J., Meir R., Maiorov V., (1996), Towards Robust Model Selection using Estimation and Approximation Error Bounds, *Proc.* 9^{th} *Annual Conference on Computational Learning Theory*, p.57, ACM, New York N.Y..

Ratsaby J., (1998), Incremental Learning with Sample Queries, *IEEE Trans. on PAMI*, Vol. 20, No. 8, Aug. 1998.

Ratsaby J., (2003), On Learning Multicategory Classification with Sample Queries,
`http://www.cs.ucl.ac.uk/staff/J.Ratsaby/Publications/PDF/m-class.pdf`.

Shawe-Taylor J., Bartlett P., Williamson R., Anthony M., (1996), A Framework for Structural Risk Minimisation. NeuroCOLT Technical Report Series, NC-TR-96-032, Royal Holloway, University of London.

Valiant L. G., A Theory of the learnable, (1984), *Comm. ACM* 27:11, p. 1134-1142.

Vapnik V.N., (1982), “Estimation of Dependences Based on Empirical Data”, Springer-Verlag, Berlin.