JMLR: Workshop and Conference Proceedings 19 (2011) 69–86 24th Annual Conference on Learning Theory

Oracle inequalities for computationally budgeted model selection

Alekh Agarwal ALEKH@CS.BERKELEY.EDU
University of California, Berkeley

John C. Duchi JDUCHI@CS.BERKELEY.EDU
University of California, Berkeley

Peter L. Bartlett BARTLETT@CS.BERKELEY.EDU
University of California, Berkeley and Queensland University of Technology

Clement Levrard CLEMENT@ENS.FR
École Normale Supérieure

Editor: Sham Kakade, Ulrike von Luxburg

Abstract

We analyze general model selection procedures using penalized empirical loss minimization under computational constraints. While classical model selection approaches do not consider computational aspects of performing model selection, we argue that any practical model selection procedure must not only trade off estimation and approximation error, but also the effects of the computational effort required to compute empirical minimizers for different function classes. We provide a framework for analyzing such problems, and we give algorithms for model selection under a computational budget. These algorithms satisfy oracle inequalities that show that the risk of the selected model is not much worse than if we had devoted all of our computational budget to the best function class.

Keywords: Model selection, oracle inequalities, computational budget

1. Introduction

In the standard statistical prediction setting, one receives samples $\{z_1, \ldots, z_n\} \subseteq \mathcal{Z}$ drawn i.i.d. from some unknown distribution P over a sample space $\mathcal{Z}$, and given a loss function ℓ, seeks a function f to minimize the risk

$$R(f) := \mathbb{E}[\ell(z, f)]. \tag{1}$$

Since $R(f)$ is unknown, the typical approach is to (approximately) minimize the empirical risk, $\widehat{R}_n(f) := \frac{1}{n}\sum_{i=1}^n \ell(z_i, f)$, over a function class $\mathcal{F}$. We seek a function f_n with a risk close to the Bayes risk, the minimal risk over all measurable functions, which is $R_0 := \inf_f R(f)$. There is a natural tradeoff based on the class $\mathcal{F}$ one chooses, since

$$R(f_n) - R_0 = \left(R(f_n) - \inf_{f\in\mathcal{F}} R(f)\right) + \left(\inf_{f\in\mathcal{F}} R(f) - R_0\right),$$

which decomposes the excess risk of f_n into estimation error (left) and approximation error (right).

A common approach to addressing this tradeoff is to express $\mathcal{F}$ as a union of classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$. The *model selection problem* is to choose a class $\mathcal{F}_i$ and a function $f \in \mathcal{F}_i$ to give the best tradeoff between estimation error and approximation error.[1] A common approach to the model selection problem is the now classical idea of *complexity regularization*, which arose out of early works by Mallows (1973) and Akaike (1974). The complexity regularization approach balances two competing objectives: the minimum empirical risk of a model class $\mathcal{F}_i$ (approximation error) and a complexity penalty (to control estimation error) for the class. Different choices of the complexity penalty give rise to different model selection criteria and algorithms (see e.g. Massart, 2003, and the references therein). Results of several authors (e.g. Bartlett et al., 2002; Lugosi and Wegkamp, 2004; Massart, 2003) show that given a dataset of size n, the output $\widehat{f}_n$ of the procedure roughly satisfies

$$\mathbb{E}R(\widehat{f}_n) - R_0 \leq \min_i \left[\inf_{f \in \mathcal{F}_i} R(f) - R_0 + \gamma_i(n) \right] + \mathcal{O}_p\left(\frac{1}{\sqrt{n}}\right), \tag{2}$$

where $\gamma_i(n)$ is a complexity penalty for class i, which is usually decreasing to zero in n and increasing in i. (Several approaches to complexity regularization are possible, and an incomplete bibliography includes Vapnik and Chervonenkis, 1974; Geman and Hwang, 1982; Rissanen, 1983; Barron, 1991; Bartlett et al., 2002; Lugosi and Wegkamp, 2004).

These oracle inequalities show that, for a given sample size, the model selection procedure gives the best trade-off between the approximation and estimation errors. A drawback with the above mentioned approaches is that we need to be able to optimize over each model in the hierarchy on the entire data, in order to prove guarantees on the result of the model selection procedure. This is natural when the sample size is the key limitation, and it is computationally feasible when the sample size is small and the samples are low-dimensional. However, the cost of training K different model classes on the entire data sequence can be prohibitive when the datasets become large and high-dimensional as is common in modern settings. In these cases, it is computational resources—rather than the sample size—that are the key constraint. In this paper, we consider model selection from this computational perspective, viewing the amount of computation, rather than the sample size, as the parameter which will enter our oracle inequalities. Specifically, we consider model selection methods that work within a given computational budget.

An interesting and difficult aspect of the problem that we must address is the interaction between model class complexity and computation time. It is natural to assume that for a fixed sample size, it is more expensive to estimate a model from a complex class than a simple class. Put inversely, given a computational bound, a simple model class can fit a model to a much larger sample size than a rich model class. So any strategy for model selection under a computational budget constraint should trade off two criteria: (i) the relative training cost of different model classes, which allows simpler classes to receive far more data (thus making them resilient to overfitting), and (ii) lower approximation error in the more complex model classes.

In addressing these computational and statistical issues, this paper makes three main contributions. First, we propose a novel computational perspective on the model selection problem, which we believe should be a natural consideration in statistical learning problems.

1. In general, the number of classes K can be infinite, though we restrict attention to finitely many classes for this paper.

Secondly, within this framework, we provide an algorithm—exploiting algorithms for multi-armed bandit problems—that uses confidence bounds based on concentration inequalities to select a good model under a given computational budget. We also prove a minimax optimal oracle inequality on the performance of the selected model. Our third main contribution is another algorithm based on a coarse-grid search, for model hierarchies that are structured by inclusion, that is, $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}_K$. Under natural assumptions regarding the growth of the complexity penalties as we go to more complex classes, the coarse-grid search procedure satisfies better oracle inequalities than the earlier bandit algorithm. Both of our algorithms are computationally simple and efficient.

The remainder of this paper is organized as follows. In the next section, we formalize our setting and present the algorithms. Section 3 presents our main results as well as some consequences for specific problems and examples. We provide proofs in Sections 4 through 5; Section 4 contains the proof of the result for unstructured model selection problems, while Sec. 5 contains the proofs of oracle inequalities for model selection problems with nested classes $\mathcal{F}_i$.

2. Setup and algorithms

In this section, we will describe our statistical and computational assumptions about the problem, giving examples of classes of problems and statistical procedures that satisfy the assumptions. We will follow this with descriptions of our algorithms, including intuitive explanations of the procedures.

2.1. Setup and Goals

Recall from the introduction that we have a collection of K model classes $\mathcal{F}_1, \ldots, \mathcal{F}_K$. Let us begin by describing our computational assumptions. First, we assume as our basic unit of measure a computational quantum; within this quantum, a model can be trained on any single class $\mathcal{F}_i$ using n_i samples. That is, we associate with each class $\mathcal{F}_i$ a number of samples $n_i \in \mathbb{N}$, where n_i is chosen so that training a model from class $\mathcal{F}_i$ on n_i examples requires the same amount of time as training a model from class $\mathcal{F}_j$ on n_j samples. We assume an overall time budget of T quanta, so that if we devote the entire computational budget to class i, we could use Tn_i samples to train a model.[2] Our high level goal is to derive algorithms that perform nearly as well as if an oracle gave the best model class i^* in advance, and we could devote the entire computational budget T to class i^*.

For the statistical assumptions in our problem, we take an approach similar to that of Bartlett et al. (2002), restricting our attention to complexity penalties based on concentration inequalities. Each of our model selection procedures uses a black-box algorithm $\mathcal{A}$ for fitting functions from the model class $\mathcal{F}_i$ to the data. We require that these algorithms be statistically well-behaved, in the sense that the empirical risk of $\mathcal{A}$'s output model $\widehat{f}$ is near the true risk of $\widehat{f}$. Recalling the definitions of $R, \widehat{R}$ from the introduction, and defining $[K] = \{1, \ldots, K\}$, we state our main concentration assumption:

2. The linearity assumption is essentially no loss of generality. In addition, several algorithms satisfy it. We can work with general non-linear scalings too, at the cost of significant notational burden which we choose to avoid here.

Assumption A *Let $\mathcal{A}(i,n) \in \mathcal{F}_i$ denote the output of algorithm $\mathcal{A}$ on a sample of n data points.*

(a) For each $i \in [K]$, there is a function γ_i and constants $\kappa_1, \kappa_2 > 0$ such that for any $n \in \mathbb{N}$,

$$\mathbb{P}\left(|\widehat{R}_n(\mathcal{A}(i,n)) - R(\mathcal{A}(i,n))| > \gamma_i(n) + \kappa_2\epsilon\right) \leq \kappa_1 \exp(-4n\epsilon^2). \tag{3}$$

(b) The output $\mathcal{A}(i,n)$ is a $\gamma_i(n)$-minimizer of $\widehat{R}_n$, that is,

$$\widehat{R}_n(\mathcal{A}(i,n)) - \inf_{f \in \mathcal{F}_i} \widehat{R}_n(f) \leq \gamma_i(n).$$

(c) The function $\gamma_i(n) \leq c_i n^{-\alpha_i}$ for some $\alpha_i > 0$.

(d) For any fixed function $f \in \mathcal{F}_i$, $\mathbb{P}(|\widehat{R}_n(f) - R(f)| > \kappa_2\epsilon) \leq \kappa_1 \exp(-4n\epsilon^2)$.

There are many classes of functions and corresponding algorithms that satisfy Assumption A. For one simple example, let $\{\mathcal{F}_i\}$ be VC-classes of functions, where each $\mathcal{F}_i$ has VC-dimension d_i, and ℓ be the hinge loss, where $\ell(z,f) = [1 - yf(x)]_+$. Assuming that $\ell(z,f) \leq B$ for all $f \in \mathcal{F}$, Dudley's entropy integral in this case gives (Dudley, 1978)

$$\gamma_i(n) = \mathcal{O}\left(\sqrt{\frac{d_i}{n}}\right) \quad \text{and} \quad \kappa_1 \leq 2,\ \kappa_2 = \mathcal{O}(B). \tag{4}$$

Similar results hold for other convex losses and problems, for example regression and density estimation problems with squared or log losses. For function classes of bounded complexity, such as VC, Sobolev, or Besov classes, penalty functions $\gamma_i(n)$ can be computed that satisfy Assumption A using many techniques; some relevant approaches include Rademacher and Gaussian complexities of the function classes $\mathcal{F}_i$, metric entropy, Dudley's entropy integral, or localization techniques (e.g. Pollard, 1984; Bartlett and Mendelson, 2002; Dudley, 1967). In many concrete cases, such as parametric models or VC classes, Assumption A(c) is satisfied with $\alpha_i = \frac{1}{2}$.

Our approach, similar to the idea of complexity regularization, is to perform a kind of penalized model selection. If we knew the true risk functional R, we could minimize a combination of the risk and complexity penalty based on the number of samples our computational budget allows for the class. In particular, given penalty functions γ_i, we define the best class in hindsight as

$$i^* := \operatorname*{argmin}_{i \in [K]} \left\{ \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(Tn_i) \right\}. \tag{5}$$

The idea is that an algorithm performing model selection—while taking into account its computational limitations—should choose the best class considering the total number of samples it could possibly have seen for the class. We note that this is also closely related to the criterion (2) minimized in the absence of a computational budget, but in the classical case it is assumed that each function class can be evaluated on an identical and fixed number of samples n.

FOREACH $i \in [K]$ query n_i examples from class $\mathcal{F}_i$
FOR $t = K + 1$ to T
  SET $n_i(t)$ to be the number of examples seen for class i at time t
  LET $i_t = \text{argmin}_{i \in [K]} \overline{R}(j, n_i(t)) - \sqrt{\frac{\log T}{n_i(t)}}$
  QUERY n_{i_t} examples for class i_t
OUTPUT $\widehat{i}$, the index of the most frequently selected class.

Algorithm 1: Multi-armed bandit algorithm for selection of best class $\widehat{i}$.

2.2. Upper-confidence bound algorithm without structure

We now turn to outlining the first of the two main scenarios analyzed in this paper. For now, we do not assume any structure relating the collection of model classes $\mathcal{F}_1, \ldots, \mathcal{F}_K$. The main idea of our algorithm in this case is to incrementally allocate our computational quota amongst the function classes, where we trade off receiving samples for classes that have good risk performance against exploring classes for which we have received few data points. We view the budgeted model selection problem as a repeated game with T rounds. At iteration t, the procedure allocates one additional quantum of computation to a (to be specified) function class i. We assume that the computational complexity of fitting a model grows linearly and incrementally with the number of samples, which means that allocating an additional quantum of training time allows the black-box training algorithm $\mathcal{A}$ to process an additional n_i samples for class $\mathcal{F}_i$. The linear growth assumption is satisfied, for instance, when the loss function ℓ is convex and the black-box learning algorithm $\mathcal{A}$ is a stochastic or online convex optimization procedure (e.g. Cesa-Bianchi and Lugosi, 2006; Nemirovski et al., 2009).

Using our previously defined notation, we now define the criterion we use in our procedure to select the class i to which we allocate a quantum. The optimistic selection criterion for class i, assuming that $\mathcal{F}_i$ has seen n samples at this point in the game, is

$$\overline{R}(i, n) = \widehat{R}_n(\mathcal{A}(i, n)) - \gamma_i(n) - \sqrt{\frac{\log K}{n}} + \gamma_i(Tn_i). \tag{6}$$

The intuition behind the definition of $\overline{R}(i, n)$ is that we would like the algorithm to choose functions f and classes i that minimize $\widehat{R}_n(f) + \gamma_i(Tn_i) \approx R(f) + \gamma_i(Tn_i)$, but the negative $\gamma_i(n)$ and $\sqrt{\log K/n}$ terms lower the criterion significantly when n is small and thus encourage initial exploration. The criterion (6) essentially combines the penalized model-selection objective used by Bartlett et al. (2002) (though we use a $\log K$ term, as we assume a finite number of classes) with an optimistic criterion similar to those used in multi-armed bandit algorithms (Auer et al., 2002). Algorithm 1 contains our bandit procedure for model selection. We run Alg. 1 for T rounds, where T is such that the entire computational budget is exhausted. Our results in Section 3.1 show that Alg. 1 satisfies our twofold goals of selecting the class i^* with high probability and outputting a function f with good risk performance.

2.3. Coarse-grid search algorithm for inclusion hierarchy

In practice, the classes $\mathcal{F}_i$ are rarely completely unrelated; perhaps the most common scenario in model selection is structural risk minimization, where the model classes $\mathcal{F}_i$ are subsets ordered in increasing complexity of a larger model space. To that end, our second main scenario involves studying computationally constrained model selection procedures under the following assumption.

Assumption B *The function classes $\mathcal{F}_i$ satisfy an inclusion hierarchy:*

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_K \tag{7}$$

One simple example satisfying above assumption is classes of functions of the form $x \mapsto f(x) = \langle \theta, x \rangle$, where each function class $\mathcal{F}$ is identified with an increasing bound on $\|\theta\|$. A second simple family of examples consists of scenarios in which $f \in \mathcal{F}_i$ is of the form $x \mapsto f(x) = \langle \theta, \phi_i(x) \rangle$ where ϕ_i is a feature mapping of the input data $\mathcal{Z}$ and ϕ_i is a projection of ϕ_{i+1}. For example, functions in class $i+1$ observe more features than those in class i or the different classes $\mathcal{F}_i$ may consist of an increasing sequence of wavelet bases.

Intuitively, we expect the structure assumed above to help our model selection procedure because the minimum expected risks of different function classes are no longer independent of each other. It is easy to see that under our assumption,

$$R_i^* \leq R_j^* \quad \text{for } i \geq j. \tag{8}$$

Clearly, under Assumption B, the penalties can always be chosen to be increasing as a function of the class complexity:

$$\gamma_i(n) \geq \gamma_j(n) \quad \text{for } i \geq j. \tag{9}$$

Since our approach involves giving a different number of samples to each class, we require a slightly stronger ordering than the above equation. We assume that for any budget T, we have

$$\gamma_i(Tn_i) \geq \gamma_j(Tn_j) \quad \text{for } i \geq j. \tag{10}$$

This assumption is reasonable since we expect that $\gamma_i(n)$ is a decreasing function of n and $n_i \leq n_j$ for $i \geq j$, so that $\gamma_i(Tn_i) \geq \gamma_i(Tn_j) \geq \gamma_j(Tn_j)$.

We now show a simple grid-search based algorithm that gives oracle inequalities depending only logarithmically on the number of classes for this inclusion hierarchy under natural conditions on the growth of the complexity penalties as a function of the class index i. The method takes inspiration from the naïve strategy that splits the budget T uniformly across the K classes and finds the class with the smallest penalized empirical risk, using Tn_i/K samples for class i. Of course, the naïve approach has the drawback that the computational budget available to each class is reduced by a factor of K, which yields very poor scaling with the number K of classes.

The key observation we exploit is that under the nesting structure (7), we do not need to find the smallest regularized empirical risk for each class. We can instead pick a small subset S of classes and perform model selection only over the classes in S, then use the inclusion assumption B to reason about the classes not in S for appropriate choices of S. With this intuition, we now define a good choice for S:

Definition 1 (Coarse grid) *For a set $S \subseteq [K]$, we say that S satisfies the* coarse grid *conditions with parameters $s \in \mathbb{N}$ and $\lambda > 0$ if $|S| = s$ and for each $i \in [K]$ there is an index $j \in S$ such that*

$$\gamma_i\left(\frac{Tn_i}{s}\right) \le \gamma_j\left(\frac{Tn_j}{s}\right) \le (1+\lambda)\gamma_i\left(\frac{Tn_i}{s}\right). \tag{11}$$

We define $s(\lambda)$ to be the size of the smallest set S satisfying condition (11), noting that $s(\lambda) \le K$. In general, for a given λ there may be no small set S satisfying Definition 1; however, we are interested in settings where a set S of size $s(\lambda) = \mathcal{O}(\log K)$ exists.

Example 1 *Let $\{\mathcal{F}_i\}$ be an increasing collection of VC-classes, say $f \in \mathcal{F}_i$ is of the form $x \mapsto f(x) = \langle \theta, \phi_i(x)\rangle$ where ϕ_i is a d_i-dimensional mapping and $\mathcal{F}_i$ has VC-dimension d_i. In this case, recalling the VC-bound (4), we know that (up to constant factors) $\gamma_i(n) \le \sqrt{d_i/n}$. Making the reasonable assumption that training time is linearly dependent on the VC dimension, we have $n_i = n_K(d_K/d_i)$ for $i \in [K]$, so*

$$\gamma_i(Tn_i) = \gamma_i\left(\frac{Tn_K d_K}{d_i}\right) \le \sqrt{\frac{d_i^2}{Tn_K d_K}} = d_i \cdot \frac{1}{\sqrt{Tn_K d_K}}.$$

Example 1 is suggestive of a pattern common to many hierarchies of function classes—including parametric and VC-classes with i indexing VC-dimension—where the penalty functions interact with the sample sizes n_i so that γ_i splits naturally into a product $\gamma_i(Tn_i) = g(T)h(i)$ for some functions g and h (which may depend on K). For such cases, the condition (11) reduces to ensuring

$$g\left(\frac{T}{s(\lambda)}\right)h(i) \le g\left(\frac{T}{s(\lambda)}\right)h(j) \le (1+\lambda)g\left(\frac{T}{s(\lambda)}\right)h(i),$$

which amounts to showing $h(i) \le h(j) \le (1+\lambda)h(i)$ independent of the setting of $s(\lambda)$ (since h is non-increasing, we need only show the latter inequality). Let $S = \{j_1, \ldots, j_{s(\lambda)}\}$. We construct S by setting $j_{s(\lambda)} = K$ and recursively defining j_i to be the smallest index $j_i < j_{i+1}$ such that

$$h(j_{i+1}) \le (1+\lambda)h(j_i).$$

Then the number of classes can be bounded by using the relation

$$h(K) = h(j_{s(\lambda)}) \le (1+\lambda)h(j_{s(\lambda)-1}) \le \cdots \le (1+\lambda)^{s(\lambda)}h(1),$$

so that so long as $s(\lambda) \ge \frac{\log(h(K)/h(1))}{\log(1+\lambda)}$, we can choose a set S satisfying condition (11) with $|S| = s(\lambda)$. In particular, $s(\lambda)$ is logarithmic in K as long as the function h grows sub-exponentially. Other natural examples of function classes satisfying such growth conditions include Besov or Sobolev function classes nested by degree or smoothness as well as wavelet bases. We refer the reader to the work of Barron et al. (1999) for a compendium of results where $\gamma_i(Tn_i) = g(T)h(i)$.

Given the above, our algorithm has a simple description. We fix a desired accuracy λ and find the smallest set S satisfying Definition 1. We then pick the class $\widehat{i}$ satisfying

$$\widehat{i} \in \operatorname*{argmin}_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i\left(Tn_i/s(\lambda)\right) \right\}, \tag{12}$$

where $|S| = s(\lambda)$. We observe that the penalty functions are typically known in closed form (with the exception of data-dependent complexity penalties), and hence computation of the set S can be efficient and is (generally) much cheaper than training the models. In Section 3.2, we give an oracle inequality on the performance of the estimate $\widehat{i}$ from the procedure (12) that has only mild dependence on the number of classes so long as $s(\lambda)$ does not grow too fast with K.

3. Main results and their consequences

In this section, we come to the description of the performance guarantees for Algorithms 1 and (12). To build intuition, we also specialize the theorems to specific statistical problems and model classes.

3.1. Oracle inequalities for unstructured model classes

In this section we give performance guarantees on the class picked by Algorithm 1. We define the excess penalized risk

$$\Delta_i := R_i^* + \gamma_i(Tn_i) - R_{i^*}^* - \gamma_{i^*}(Tn_{i^*}) \geq 0. \tag{13}$$

Essentially without loss of generality, we assume that the infimum in the equation $R_i^* = \inf_{f \in \mathcal{F}_i} R(f)$ is attained by a function f_i^*. If the infimum is not attained we simply choose some fixed f_i^* such that $R(f_i^*) \leq \inf_{f \in \mathcal{F}_i} R(f) + \delta$ for an arbitrarily small $\delta > 0$. We first perform analysis under the assumption that $\Delta_i > 0$ strictly for $i \neq i^*$, but we will then relax to allow non-unique i^*.

The gains of a computationally adaptive strategy over naïve strategies are best seen when the gap (13) is non-zero. Under this assumption, we can follow the ideas of Auer et al. (2002) and show that the fraction of the computational budget allocated to any suboptimal class $i \neq i^*$ goes quickly to zero as T grows. We provide the proof of the following theorem in Section 4.

Theorem 2 *Let Alg. 1 be run for T rounds and $T_i(T)$ be the number of times class i is queried. Let Δ_i be defined as in (13), the conditions of Assumption A hold, and assume that $T \geq K$. Define $\beta_i = \max\{1/\alpha_i, 2\}$. There is a constant C such that*

$$\mathbb{E}[T_i(T)] \leq \frac{C}{n_i}\left(\frac{c_i + \kappa_2\sqrt{\log T}}{\Delta_i}\right)^{\beta_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > \frac{C}{n_i}\left(\frac{c_i + \kappa_2\sqrt{\log T}}{\Delta_i}\right)^{\beta_i}\right) \leq \frac{\kappa_1}{TK^4},$$

where c_i and α_i come from the definition of the concentration function γ_i in Assumption A(c).

At a high level, this result shows that the fraction of budget allocated to any suboptimal class goes to 0 at the rate $\frac{1}{n_i T}\left(\frac{\sqrt{\log T}}{\Delta_i}\right)^{\beta_i}$. Hence, asymptotically in T, we will receive exponentially more samples for i^* than any other class and will perform almost as well as if we had known i^* in advance. To see an example of concrete rates that can be concluded from the above result, let $\mathcal{F}_1, \ldots, \mathcal{F}_K$ be model classes with finite VC-dimension,[3] so that Assumption A is satisfied with $\alpha_i = \frac{1}{2}$. Then we have

3. Similar corollaries hold for any model class whose metric entropy grows as polylog $\left(\frac{1}{\epsilon}\right)$.

Corollary 3 *Under the conditions of Theorem 2, assume $\mathcal{F}_1, \ldots, \mathcal{F}_K$ are model classes of finite VC-dimension, where $\mathcal{F}_i$ has dimension d_i. Then there is a constant C such that*

$$\mathbb{E}[T_i(T)] \leq C\frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i} \quad \text{and} \quad \mathbb{P}\left(T_i(T) > C\frac{\max\{d_i, \kappa_2^2 \log T\}}{\Delta_i^2 n_i}\right) \leq \frac{\kappa_1}{TK^4}.$$

The result of Corollary 3 is nearly optimal in general due to a lower bound for the special case of multi-armed bandit problems (Lai and Robbins, 1985). To see the connection, let $\mathcal{F}_i$ correspond to the ith arm in a multi-armed bandit problem and the risk R_i^* be the expected reward of arm i. In this case, the complexity penalty γ_i for each class is 0. Lai and Robbins give a lower bound that shows that the expected number of pulls of any suboptimal arm is at least $\mathbb{E}[T_i(T)] = \Omega\left(\frac{\log T}{\mathrm{KL}(p_i \| p_{i^*})}\right)$, where p_i and p_{i^*} are the reward distributions for the ith and optimal arms, respectively.

Unfortunately, the condition that $\Delta_i > 0$ may not always be satisfied, or Δ_i may be so small as to render the bound in Theorem 2 vacuous. Nevertheless, we intuitively believe that our algorithm can quickly find a small set of "good" classes—those with small penalized risk—and spend its computational budget to try to distinguish amongst them. In this case, though, Algorithm 1 will not visit suboptimal classes and so can still output a function f satisfying good oracle bounds. In order to prove a result quantifying this intuition, we first upper bound the *regret* of Algorithm 1, that is, the average excess risk suffered by the algorithm over all iterations, and then show how to use this bound for obtaining a model with a small risk. We state our results for the case where $\alpha_i \equiv \alpha$ and define $\beta = \max\{1/\alpha, 2\}$.

Proposition 4 *Use the same assumptions as Theorem 2, but further assume that $\alpha_i \equiv \alpha$ for all i. With probability at least $1 - \kappa_1/TK^3$, the regret (average excess risk) of Algorithm 1 is bounded as*

$$\sum_{i=1}^{K} \Delta_i T_i(T) \leq 2eT^{1-1/\beta}\left(C\sum_{i=1}^{K}\frac{(c_i + \kappa_2\sqrt{\log T})^\beta}{n_i}\right)^{1/\beta}$$

for a constant C dependent on α.

In order to obtain a model with a small risk, we need to make an additional assumption that the models are compatible in the sense that one can define the addition operator $f + g$ for $f \in \mathcal{F}_i, g \in \mathcal{F}_j$ meaningfully. We also assume that the risk functional $R(f)$ is convex in f. In such a setting, we can average the functions minimizing the objective $\overline{R}(i, n)$, that is, $f_t = \mathrm{argmin}_{f \in \mathcal{F}_{i_t}} \widehat{R}_{n_i(t)}(f)$, to obtain a function satisfying the desired oracle inequality. For this theorem, we also assume that the constants c_i from Assumption A(c) satisfy $c_i = \mathcal{O}(\sqrt{\log T})$.

Theorem 5 *Use the same assumptions as Proposition 4. Let f_t be the function chosen by algorithm $\mathcal{A}$ at round t of Alg. 1 and define the average function $\widehat{f}_T = \frac{1}{T}\sum_{t=1}^{T} f_t$. If the risk functional R is convex, there are constants C, C' (dependent on α) such that with*

probability greater than $1 - 2\kappa_2/(TK^3)$,

$$R(\widehat{f}_T) \leq R^* + \gamma_{i^*}(Tn_{i^*}) + 2e\kappa_2 T^{-\beta}\sqrt{\log T}\left(\sum_{i=1}^{K}\frac{C}{n_i}\right)^{1/\beta} + C'\,T^{-1/\beta}\left(\sum_{i=1}^{K}\left[c_i n_i^{-\alpha} + \kappa_2 n_i^{-\frac{1}{2}}\sqrt{\log K} + \kappa_2 n_i^{-\frac{1}{2}}\sqrt{\log T}\right]^{\beta}\right)^{1/\beta}.$$

Let us interpret the above bound and discuss its optimality. When $\alpha = \frac{1}{2}$ (e.g., for VC classes), we have $\beta = 2$; moreover, it is clear that $\sum_{i=1}^{K}\frac{C}{n_i} = \mathcal{O}(K)$. Thus, to within constant factors, we have

$$R(\widehat{f}_T) = R^* + \gamma_{i^*}(Tn_{i^*}) + \mathcal{O}\left(\frac{\sqrt{K\max\{\log T, \log K\}}}{\sqrt{T}}\right).$$

Ignoring logarithmic factors, the above bound is minimax optimal, which follows by a reduction of our model selection problem to the special case of a multi-armed bandit problem. In this case, Theorem 5.1 of Auer et al. (2003) shows that for any set of K, T values, there is a distribution over the rewards of arms which forces $\Omega(\sqrt{KT})$ regret, that is, the average excess risk of the classes chosen by Alg. 1 must be $\Omega(\sqrt{KT})$. We provide proofs of Proposition 4 and Theorem 5 in the long version of the paper.

3.2. Oracle inequalities for nested hierarchies

In this section we provide an oracle inequality on the output of the procedure (12) that has a more favorable dependence on the number of classes K than our bounds for unstructured function classes $\mathcal{F}_i$. The main idea is to use Assumption B along with Definition 1 to show that performing a coarse grid search over S is sufficient to deduce an oracle inequality over the entire hierarchy. The next theorem provides an oracle inequality for the risk of the function $f = \mathcal{A}(\widehat{i}, n_{\widehat{i}}T/s(\lambda))$, which is the output of the learning algorithm $\mathcal{A}$ applied to the class $\widehat{i}$ picked by our algorithm.

Theorem 6 *Let $f = \mathcal{A}(\widehat{i}, n_{\widehat{i}}T/s(\lambda))$ be the output of the algorithm $\mathcal{A}$ for class $\widehat{i}$ specified by the procedure (12). Let Assumptions A–B hold. With probability at least $1 - 3\kappa_1\exp(-4m)$*

$$R(f) \leq \min_{i\in[K]}\left\{R_i^* + 2(1+\lambda)\gamma_i\left(\frac{Tn_i}{s(\lambda)}\right)\right\} + \kappa_2\sqrt{\frac{s(\lambda)\log K}{2Tn_K}} + \kappa_2\sqrt{\frac{ms(\lambda)}{Tn_K}}.$$

Remark: It is possible to reduce the $\kappa_2\sqrt{s(\lambda)\log K/2Tn_K}$ term in the bound above to a $(1+\lambda)\sqrt{s(\lambda)\log i/2Tn_i}$ term appearing inside the minimum over classes $i \in [K]$ by requiring the coarse grid condition (11) to hold over terms of the form $\gamma_i(Tn_i/s(\lambda)) + \sqrt{s(\lambda)\log i/2Tn_i}$. This stronger bound applies, for example, to sequences of VC-classes as described in Example 1.

The above result makes it clear that the excess risk of the algorithm—outside of the minimum over all the classes—scales as $\mathcal{O}(T^{-1/2})$. It is of interest to contrast Theorem 6

with results of the previous section. In the completely general case, we have a dependence on K better than $\sqrt{K}$ only when there is constant separation between the penalized risks of different classes. Since $s(\lambda) \leq K$, the result of the above theorem is essentially as strong as any of the results from the previous section, as we would hope when we know $\mathcal{F}_i \subseteq \mathcal{F}_{i+1}$.

Nonetheless, the main strength of Theorem 6 is in scenarios where $s(\lambda) = \mathcal{O}(\log K)$, such as VC-classes (e.g. Example 1) with at most polynomial growth in VC-dimension. In such scenarios, the function f that the procedure outputs is competitive (up to logarithmic factors) with an oracle that devotes the entire computation budget to the optimal class. We note that model selection procedures suffer a penalty of $\sqrt{\log K}$ (or $\sqrt{\log i}$) even in computationally unconstrained settings (see, e.g., Bartlett et al., 2002), so our computationally restricted procedure suffers at most an additional penalty of $\mathcal{O}(\sqrt{\log K})$. We conclude by recalling that many common model selection scenarios satisfy $s(\lambda) = \mathcal{O}(\log K)$, as noted in Section 2.3.

4. Proof of Theorem 2

At a high level, the proof of this theorem involves combining the techniques for analysis of multi-armed bandits developed by Auer et al. (2002) with Assumption A. We start by giving a lemma which will be useful to prove the theorem. The lemma states that after a sufficient number of initial iterations τ, the probability Algorithm 1 chooses to receive samples for a sub-optimal function class $i \neq i^*$ is extremely small. Recall also our notational convention that $\beta_i = \max\{1/\alpha_i, 2\}$.

Lemma 7 *For any class i, any $s_i \in [1, T]$ and $s_{i^*} \in [\tau, T]$ where $\tau > 0$ satisfies*

$$\tau > \frac{2^{\beta_i}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{\beta_i}}{n_i \Delta_i^{\beta_i}},$$

under Assumption A we have

$$\mathbb{P}\left(\overline{R}(i, n_i s_i) - \kappa_2\sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2\sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right) \leq \frac{2\kappa_1}{(TK)^4}.$$

We defer the proof of the lemma to Appendix A, though at a high level the proof works as follows. The "bad event" in Lemma 7, that is, Algorithm 1 selects a sub-optimal class $i \neq i^*$, occurs only if one of the following three errors occurs: the empirical risk of class i is much lower than its true risk, the empirical risk of class i^* is higher than its true risk, or s_i is not large enough to actually separate the true penalized risks from one another. Under the assumptions of the lemma, however, coupled with the uniform convergence properties in Assumption A, each of these three sub-events is quite unlikely. Now we turn to the proof of Theorem 2 assuming the lemma.

Let i_t denote the model class index i chosen by Algorithm 1 at time t, and let $s_i(t)$ denote the number of times class i has been selected at round t of the algorithm. When no time index is needed, s_i will denote the same thing. Note that if $i_t = i$ and the number of

times class i is queried exceeds $\tau > 0$, then by the definition of the selection criterion (6) and choice of i_t in Alg. 1, for some $s_i \in \{\tau, \ldots, t-1\}$ and $s_{i^*} \in \{1, \ldots, t-1\}$ we have

$$\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}.$$

Here we interpret $\overline{R}(i, n_i s_i)$ to mean a random realization of the observed risk consistent with the samples we observe. Using the above implication, we thus have

$$\begin{aligned} T_i(n) = 1 + \sum_{t=K+1}^{T} \mathbb{I}\left(i_t = i\right) \quad &\leq \quad \tau + \sum_{t=K+1}^{T} \mathbb{I}\left(i_t = i, T_i(t-1) \geq \tau\right) \\ &\leq \tau + \sum_{t=K+1}^{T} \mathbb{I}\left(\min_{\tau \leq s_i < t} \overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \max_{0 < s < t} \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right) \\ &\leq \tau + \sum_{t=1}^{T} \sum_{s_{i^*}=1}^{t-1} \sum_{s_i=\tau}^{t-1} \mathbb{I}\left(\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right). \quad (14) \end{aligned}$$

To control the last term, we invoke Lemma 7 and obtain that

$$\tau > \frac{2^{\beta_i}(c_i + \kappa_2 \sqrt{\log T} + \kappa_2 \sqrt{\log K})^{\beta_i}}{n_i \Delta_i^{\beta_i}} \quad \Rightarrow \quad \mathbb{E}[T_i(n)] \leq \tau + \sum_{t=1}^{T} \sum_{s=1}^{t-1} \sum_{s_i=\tau}^{t-1} 2 \frac{\kappa_1}{(TK)^4} \leq \tau + \frac{\kappa_1}{TK^4}.$$

Hence for any suboptimal class $i \neq i^*$, $\mathbb{E}[T_i(n)] \leq \tau_i + \kappa_1/(TK^4)$, where τ_i satisfies the lower bound of Lemma 7 and is thus logarithmic in T. Under the assumption that $T \geq K$, for $i \neq i^*$,

$$\mathbb{E}[T_i(T)] \leq C \frac{(c_i + \kappa_2 \sqrt{\log T})^{\max\{1/\alpha_i, 2\}}}{n_i \Delta_i^{\max\{1/\alpha_i, 2\}}} \quad (15)$$

for a constant $C \leq 2 \cdot 4^{\max\{1/\alpha_i, 2\}}$. Now we prove the high-probability bound. For this part, we need only concern ourselves with the sum of indicators from (14). Markov's inequality shows that

$$\mathbb{P}\left(\sum_{t=K+1}^{T} \mathbb{I}\left(i_t = i, T_i(t-1) \geq \tau\right) \geq 1\right) \leq \frac{\kappa_1}{TK^4}.$$

Thus we can assert that the bound (15) on $T_i(T)$ holds with high probability.

Remark: By examining the proof of Theorem 2, it is straightforward to see that if we modify the multipliers on the $\sqrt{\cdot}$ terms in the criterion (6) by $m\kappa_2$ instead of κ_2, we get that the probability bound is of the order $T^{3-4m^2} K^{-4m^2}$, while the bound on $T_i(T)$ is scaled by m^{1/α_i}.

5. Model selection over nested hierarchies

In this section, we prove Theorem 6. The following proposition states that the class returned by the output of the procedure (12) satisfies an oracle inequality over the set S.

Proposition 8 *Let $f = \mathcal{A}(\widehat{i}, n_{\widehat{i}}T/s(\lambda))$ be the output of the algorithm $\mathcal{A}$ for class $\widehat{i}$ specified in Equation 12. Under the conditions of Theorem 6, with probability at least $1 - 3\kappa_1 \exp(-4m)$*

$$R(f) \leq \min_{i \in S} \left\{ R_i^* + \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}}.$$

The proof of the proposition follows from an argument similar to that given by Bartlett et al. (2002). We present a proof at the end of this section, since our setting is slightly different: each class receives a different number of independent samples. First, however, we complete the proof of Theorem 6 using the proposition.

Proof of Theorem 6 Let $i \in [K]$ be any class (not necessarily in S), and let $j \in S$ be the smallest class satisfying $j \geq i$. Then by construction of S, we know that

$$\gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) \leq \gamma_j \left(\frac{Tn_j}{s(\lambda)} \right) \leq (1 + \lambda) \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right).$$

Thus we can lower bound the penalized risk of class i as

$$R_i^* + 2(1 + \lambda) \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) \geq R_j^* + 2\gamma_j \left(\frac{Tn_j}{s(\lambda)} \right),$$

where we used the nesting assumption B to conclude that $j \geq i$ implies $R_j^* \leq R_i^*$.

Now combining the above lower bound with the inequality in Proposition 8 yields that with probability at least $1 - 3\kappa_1 \exp(-m)$

$$\begin{aligned} R(f) &\leq \min_{i \in S} \left\{ R_i^* + \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \sqrt{\frac{ms(\lambda)}{Tn_K}} \\ &\leq \min_{i \in [K]} \left\{ R_i^* + 2(1 + \lambda) \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) \right\} + \kappa_2 \sqrt{\frac{s(\lambda) \log K}{2Tn_K}} + \kappa_2 \sqrt{\frac{ms(\lambda)}{Tn_K}} \end{aligned}$$

since $K \geq i$ and $n_i \geq n_K$. ■

Proof of Proposition 8 To prove the proposition, we would like to control the probability

$$\begin{aligned} &\mathbb{P} \left[R(f) > \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \epsilon \right] \\ &\leq \underbrace{\mathbb{P} \left[R(f) > \min_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) \right\} + \epsilon/2 \right]}_{\mathcal{T}_1} \qquad (16) \\ &\quad + \underbrace{\mathbb{P} \left[\min_{i \in S} \left\{ \widehat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) \right\} > \min_{i \in S} \left\{ R_i^* + 2\gamma_i \left(\frac{Tn_i}{s(\lambda)} \right) + \kappa_2 \sqrt{\frac{s(\lambda) \log i}{2Tn_i}} \right\} + \epsilon/2 \right]}_{\mathcal{T}_2} \end{aligned}$$

where the inequality follows from a union bound.

We now bound the terms $\mathcal{T}_1$ and $\mathcal{T}_2$ separately. To bound the terms, we first observe that by the construction (12), the minimum over the penalized empirical risk is attained for the class $\widehat{i}$. We thus simplify $\mathcal{T}_1$ as

$$\mathbb{P}\left[R(f) > \min_{i\in S}\left\{\widehat{R}_{Tn_i/s(\lambda)}(i) + \gamma_i\left(\frac{Tn_i}{s(\lambda)}\right)\right\} + \epsilon/2\right] = \mathbb{P}\left[R(f) > \left\{\widehat{R}(f) + \gamma_{\widehat{i}}\left(\frac{Tn_{\widehat{i}}}{s(\lambda)}\right)\right\} + \epsilon/2\right]$$
$$\leq \kappa_1 \exp\left(-\frac{Tn_{\widehat{i}}\epsilon^2}{\kappa_2^2 s(\lambda)}\right),$$

where the inequality follows by application of Assumption A(a). To bound $\mathcal{T}_2$ in the sum (16), we define $f_i^* = \operatorname{argmin}_{f\in\mathcal{F}_i} R(f)$ so that $R_i^* = R(f)$. Noting that the event in $\mathcal{T}_2$ implies that

$$\max_{i\in S}\left\{\widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i\left(\frac{Tn_i}{s(\lambda)}\right) - \kappa_2\sqrt{\frac{s(\lambda)\log i}{2Tn_i}}\right\} > \frac{\epsilon}{2},$$

we can use the union bound to see

$$\mathcal{T}_2 \leq \mathbb{P}\left[\sup_{i\in S}\left\{\widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i\left(\frac{Tn_i}{s(\lambda)}\right) - \kappa_2\sqrt{\frac{s(\lambda)\log i}{2Tn_i}}\right\} > \frac{\epsilon}{2}\right]$$
$$\leq \sum_{i\in S}\mathbb{P}\left[\widehat{R}_{Tn_i/s(\lambda)}(i) - R_i^* - \gamma_i\left(\frac{Tn_i}{s(\lambda)}\right) - \kappa_2\sqrt{\frac{s(\lambda)\log i}{2Tn_i}} > \frac{\epsilon}{2}\right]$$
$$\leq \sum_{i\in S}\mathbb{P}\left[\widehat{R}(f_i^*) - R_i^* > \frac{\epsilon}{2} + \kappa_2\sqrt{\frac{s(\lambda)\log i}{2Tn_i}}\right],$$

where the final inequality uses Assumption A(b), which states that $\mathcal{A}$ outputs a γ_i-minimizer of the empirical risk. Now we can bound the deviations using Assumption A(d), since f_i^* is non-random:

$$\mathcal{T}_2 \leq \sum_{i\in S}\kappa_1 \exp\left(-\frac{Tn_i\epsilon^2}{s(\lambda)\kappa_2^2}\right)\exp\left(-2\log i\right).$$

Setting $\epsilon = \kappa_2\sqrt{\frac{ms(\lambda)}{Tn_K}}$, see that the first term in bounding $\mathcal{T}_2$ reduces to $\exp(-mn_i/n_K) \leq \exp(-m)$ since $n_i \geq n_K$. Then we get

$$\mathcal{T}_2 \leq \sum_{i\in S}\kappa_1\exp(-m)\exp\left(-2\log i\right)$$
$$\leq 2\kappa_1\exp(-m),$$

where the last step uses $\sum_{i=1}^{\infty} 1/i^2 = \pi^2/6 \leq 2$. Finally, plugging the stated setting of ϵ into the bound on $\mathcal{T}_1$ completes the proof. ■

Acknowledgments

In performing this research, AA was supported by a Microsoft Research Fellowship, and JCD was supported by the National Defense Science and Engineering Graduate Fellowship (NDSEG) Program. AA and PB gratefully acknowledge the support of the NSF under award DMS-0830410.

References

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

P. Auer, N. Cesa-Bianchi, Y. Freund, and R. E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM J. Comput.*, 32(1):48–77, 2003.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Mach. Learn.*, 47(2-3):235–256, 2002. ISSN 0885-6125.

A. R. Barron. Complexity regularization with application to artificial neural networks. In *Nonparametric functional estimation and related topics*, pages 561–576. Kluwer Academic, 1991.

Andrew Barron, Lucien Birgé, and Pascal Massart. Risk bounds for model selection via penalization. *Probability Theory and Related Fields*, 113:301–413, 1999.

P. L. Bartlett and S. Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

P. L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

R. M. Dudley. The sizes of compact subsets of Hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1:290–330, 1967.

R. M. Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, 6 (6):899–929, 1978.

S. Geman and C. R. Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *Annals of Statistics*, 10:401–414, 1982.

T. L. Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6:4–22, 1985.

G. Lugosi and M. Wegkamp. Complexity regularization via localized random penalties. *Annals of Statistics*, 32(4):1679–1697, 2004.

C. L. Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

P. Massart. Concentration inequalities and model selection. In J. Picard, editor, *Ecole d'Et de Probabilits de Saint-Flour XXXIII - 2003 Series*. Springer, 2003.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

Jorma Rissanen. A universal prior for integers and estimation by minimum description length. *The Annals of Statistics*, 11(2):416–431, 1983.

V. N. Vapnik and A. Ya. Chervonenkis. *Theory of pattern recognition.* Nauka, Moscow, 1974. (In Russian).

Appendix A. Proof of Lemma 7

Following Auer et al. (2002), we show that the event in the lemma occurs with very low probability by breaking it up into smaller events more amenable to analysis. Recall that we're interested in controlling the probability of the event

$$\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \leq \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \tag{17}$$

For this bad event to happen, at least one of the following three events must happen:

$$\widehat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) - \inf_{f \in \mathcal{F}_i} R(f) \leq -\gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \tag{18a}$$

$$\widehat{R}_{n_{i^*} s_{i^*}}(\mathcal{A}(i^*, n_{i^*} s_{i^*})) - \inf_{f \in \mathcal{F}_{i^*}} R(f) \geq \gamma_i(n_{i^*} s_{i^*}) + \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2 \sqrt{\frac{\log T}{n_{i^*} s_{i^*}}} \tag{18b}$$

$$R_i^* + \gamma_i(T n_i) \leq R^* + \gamma_{i^*}(T n_{i^*}) + 2\left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i s_i}}\right). \tag{18c}$$

Temporarily use the shorthand $f_i = \mathcal{A}(i, n_i s_i)$ and $f_{i^*} = \mathcal{A}(i^*, n_{i^*} s_{i^*})$. The relationship between Eqs. (18a)–(18c) and the event in (17) follows from the fact that if none of (18a)–(18c) occur, then

$$\begin{aligned}
&\overline{R}(i, n_i s_i) - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\
&\quad = \quad \widehat{R}_{n_i s_i}(f_i) + \gamma_i(T n_i) - \gamma_i(n_i s_i) - \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} - \kappa_2 \sqrt{\frac{\log T}{n_i s_i}} \\
&\quad \overset{(18a)}{>} \quad \inf_{f \in \mathcal{F}_i} R(f) + \gamma_i(T n_i) - 2\left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log t}{n_i s_i}}\right) \\
&\quad \overset{(18c)}{>} \quad \inf_{f \in \mathcal{F}_{i^*}} R(f) + \gamma_{i^*}(T n_{i^*}) + 2\left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log T}{n_i s_i}}\right) \\
&\qquad\qquad - 2\left(\gamma_i(n_i s_i) + \kappa_2 \sqrt{\frac{\log K}{n_i s_i}} + \kappa_2 \sqrt{\frac{\log n}{n_i s_i}}\right) \\
&\quad \overset{(18b)}{>} \quad \widehat{R}_{n_{i^*} s_{i^*}}(f_{i^*}) + \gamma_{i^*}(T n_{i^*}) - \gamma_i(n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} - \kappa_2 \sqrt{\frac{\log t}{n_{i^*} s_{i^*}}} \\
&\quad = \quad \overline{R}(i^*, n_{i^*} s_{i^*}) - \kappa_2 \sqrt{\frac{\log t}{n_{i^*} s_{i^*}}}.
\end{aligned}$$

From the above string of inequalities, to show that the event (17) has low probability, we need simply show that each of (18a), (18b), and (18c) have low probability.

To prove that each of the bad events have low probability, we note the following consequences of Assumption A. Recall the definition of f_i^* as the minimizer of $R(f)$ over the class $\mathcal{F}_i$. Then by Assumption A(a),

$$R(f_i^*) - \gamma_i(n) - \kappa_2\epsilon \leq R(\mathcal{A}(i,n)) - \gamma_i(n) - \kappa_2\epsilon < \widehat{R}_n(\mathcal{A}(i,n)),$$

while Assumptions A(b) and A(d) imply

$$\widehat{R}_n(\mathcal{A}(i,n)) \leq \widehat{R}_n(f_i^*) + \gamma_i(n) \leq R(f_i^*) + \gamma_i(n) + \kappa_2\epsilon,$$

each with probability at least $1 - \kappa_1 \exp(-4n\epsilon^2)$. In particular, we see that the events (18a) and (18b) have low probability:

$$\begin{aligned}
&\mathbb{P}\left[\widehat{R}_{n_i s_i}(\mathcal{A}(i, n_i s_i)) - R(f_i^*) \leq -\gamma_i(n_i s_i) - \kappa_2\sqrt{\frac{\log K}{n_i s_i}} - \kappa_2\sqrt{\frac{\log T}{n_i s_i}}\right] \\
&\qquad \leq \kappa_1 \exp\left(-4n_i s_i\left(\frac{\log K}{n_i s_i} + \frac{\log t}{n_i s_i}\right)\right) = \frac{\kappa_1}{(tK)^4} \\
&\mathbb{P}\left[\widehat{R}_{n_{i^*} s_{i^*}}(\mathcal{A}(i^*, n_{i^*} s_{i^*})) - R^* \geq \gamma_{i^*}(n_{i^*} s_{i^*}) + \kappa_2\sqrt{\frac{\log K}{n_{i^*} s_{i^*}}} + \kappa_2\sqrt{\frac{\log T}{n_{i^*} s_{i^*}}}\right] \\
&\qquad \leq \kappa_1 \exp\left(-4n_{i^*} s_{i^*}\left(\frac{\log K}{n_{i^*} s_{i^*}} + \frac{\log T}{n_{i^*} s_{i^*}}\right)\right) = \frac{\kappa_1}{(tK)^4}.
\end{aligned}$$

What remains is to show that for large enough τ, (18c) does not happen. Recalling the definition that $R^* + \gamma_{i^*}(Tn_{i^*}) = R_i^* + \gamma_i(Tn_i) - \Delta_i$, we see that for (18c) to fail it is sufficient that

$$\Delta_i > 2\gamma_i(\tau n_i) + 2\kappa_2\sqrt{\frac{\log K}{n_i\tau}} + 2\kappa_2\sqrt{\frac{\log T}{n_i\tau}}.$$

Let $x \wedge y := \min\{x, y\}$ and $x \vee y := \max\{x, y\}$. Since $\gamma_i(n) \leq c_i n^{-\alpha_i}$, the above is satisfied when

$$\frac{\Delta_i}{2} > c_i(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} + \kappa_2\sqrt{\log K}(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} + \kappa_2\sqrt{\log T}(\tau n_i)^{-(\alpha_i \wedge \frac{1}{2})} \tag{19}$$

We can solve (19) above and see immediately that if

$$\tau_i > \frac{2^{1/\alpha_i \vee 2}(c_i + \kappa_2\sqrt{\log T} + \kappa_2\sqrt{\log K})^{1/\alpha_i \vee 2}}{n_i \Delta_i^{1/\alpha_i \vee 2}},$$

then

$$R_i^* > R^* + 2\left(\gamma_i(n_i\tau_i) + \kappa_2\sqrt{\frac{\log K}{n_i\tau_i}} + \kappa_2\sqrt{\frac{\log T}{n_i\tau_i}}\right). \tag{20}$$

Thus the event in (18c) fails to occur, completing the proof of the lemma.