Fast Rates for Estimation Error and Oracle Inequalities for Model Selection

Peter L. Bartlett

Computer Science Division and Department of Statistics

University of California, Berkeley

bartlett@cs.berkeley.edu

July 9, 2007

Abstract

We consider complexity penalization methods for model selection. These methods aim to choose a model to optimally trade off estimation and approximation errors by minimizing the sum of an empirical risk term and a complexity penalty. It is well known that if we use a bound on the maximal deviation between empirical and true risks as a complexity penalty, then the risk of our choice is no more than the approximation error plus twice the complexity penalty. There are many

cases, however, where complexity penalties like this give loose upper bounds on the estimation error. In particular, if we choose a function from a suitably simple convex function class with a strictly convex loss function, then the estimation error (the difference between the risk of the empirical risk minimizer and the minimal risk in the class) approaches zero at a faster rate than the maximal deviation between empirical and true risks. In this note, we address the question of whether it is possible to design a complexity penalized model selection method for these situations. We show that, provided the sequence of models is ordered by inclusion, in these cases we can use tight upper bounds on estimation error as a complexity penalty. Surprisingly, this is the case even in situations when the difference between the empirical risk and true risk (and indeed the error of any estimate of the approximation error) decreases much more slowly than the complexity penalty. We give an oracle inequality showing that the resulting model selection method chooses a function with risk no more than the approximation error plus a constant times the complexity penalty.

1 Introduction

Consider the following prediction problem. We have independent and identically distributed random pairs $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ from $\mathcal{X} \times \mathcal{Y}$,

where the label space $\mathcal{Y}$ is a subset of $\mathbb{R}$. The aim is to use the data $(X_1, Y_1), \ldots, (X_n, Y_n)$ to choose a function $f_n : \mathcal{X} \to \mathcal{Y}$ with small risk, $R(f_n) = \mathbf{E}\ell(Y, f_n(X))$, where ℓ is a non-negative *loss function*. Ideally, the risk should be close to the minimal or Bayes risk, $R^* = \inf_f R(f)$, where the infimum is over all measurable functions. A natural approach is to minimize the empirical risk, $\hat{R}(f) = n^{-1} \sum_{i=1}^n \ell(Y_i, f(X_i))$, over some class $\mathcal{F}$ of functions. Clearly, there is a trade-off in the choice of $\mathcal{F}$: we can decompose the excess risk of f_n as

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

where the first term represents the estimation error and the second the approximation error. We would like both terms in the decomposition to be small. For the first term to be small, $\mathcal{F}$ should be sufficiently small that it is possible to use the data to choose a near-optimal function from $\mathcal{F}$ based on the data. For the second term to be small, $\mathcal{F}$ should be sufficiently large that it contains a good approximation to the optimal prediction.

One approach to this problem is to choose a large class $\mathcal{F}$ and decompose it into

$$\mathcal{F} = \bigcup_{j \geq 1} \mathcal{F}_j,$$

then choose $\hat{f}_j \in \mathcal{F}_j$ to minimize the empirical risk, and finally choose the model index j so that $\hat{f}_j$ best balances these conflicting requirements. In

this note, we consider complexity penalization approaches, in which we choose $f_n = \hat{f}_{\hat{m}}$ with $\hat{m}$ chosen to minimize a combination of the empirical risk of $\hat{f}_j$ and a penalty term:

$$\hat{m} = \arg\min_j \left(\hat{R}(\hat{f}_j) + p_j \right),$$

where the penalty p_j depends on the sample size n.

We are interested in *oracle inequalities* of the form

$$R(f_n) - R^* \leq \inf_j \left(\inf_{f \in \mathcal{F}_j} R(f) - R^* + cp_j \right),$$

where c is a positive constant. In such an inequality, if the term cp_j decreases with n at the same rate as the estimation error $R(\hat{f}_j) - \inf_{f \in \mathcal{F}_j} R(f)$, then the inequality shows that our choice f_n has excess risk that decreases at the same rate as if we had the advice of an oracle who tells us the best complexity class $\mathcal{F}_j$ to choose. The utility of an oracle inequality of this kind depends on the accuracy with which the penalty term approximates the estimation error.

It is well known that such inequalities follow easily from uniform convergence results, as illustrated by the following theorem. The proof is immediate (see, for example, Bartlett et al., 2002).

Theorem 1. *If the data is such that*

$$\sup_j \left(\sup_{f \in \mathcal{F}_j} \left| R(f) - \hat{R}(f) \right| - p_j \right) \leq 0,$$

then in that case

$$R(f_n) - R^* \leq \inf_j \inf_{f \in \mathcal{F}_j} \left(R(f) - R^* + 2p_j \right).$$

Notice that the first inequality of the theorem involves random variables. Thus, when it holds with high probability, the second inequality of the theorem holds with high probability.

The theorem shows that it suffices to choose the penalty p_j as a high-probability upper bound on the maximal deviation between empirical risks and risks. But it is sometimes possible to obtain faster rates of convergence (smaller values of p_j as a function of n) than could be implied by the uniform convergence results. Indeed, the following theorem shows that for a nontrivial function class and loss function, the maximal deviation between $\hat{R}(f)$ and $R(f)$ can approach zero at a rate no faster than $n^{-1/2}$ (see, for example, Theorem 2.3 in Bartlett and Mendelson, 2006).

Theorem 2. *There is a constant c such that, for any loss function $\ell : \mathcal{Y}^2 \to [0,1]$ and any function class $\mathcal{F}$ for which $\sigma^2 = \sup_{f \in \mathcal{F}} \operatorname{var}(\ell(Y, f(X)) > 0$, we have*

$$\mathbf{E} \sup_{f \in \mathcal{F}} \left| \hat{R}(f) - R(f) \right| \geq \frac{c\sigma}{\sqrt{n}}.$$

On the other hand, for the quadratic loss $\ell(y, \hat{y}) = (y - \hat{y})^2$ or the exponential loss $\ell(y, \hat{y}) = \exp(-y\hat{y})$ used by AdaBoost, for example, and for $\mathcal{F}_j$ a convex bounded subset of a finite-dimensional linear class, the empirical

minimizer $\hat{f}_j$ can approach $\inf_{f \in \mathcal{F}_j} R(f)$ at the faster rate of $O(n^{-1} \log n)$, as the following theorem shows. (The theorem is an easy consequence of Theorem 3.3 in (Bartlett et al., 2005), Lemmas 14 and 15 in (Bartlett et al., 2006) and an application of the entropy integral of Dudley (1999)—see, for example the proof of Corollary 3.7 in (Bartlett et al., 2005).)

Theorem 3. *Suppose* $\ell : [0,1]^2 \to [0,1]$ *satisfies the Lipschitz condition*

$$\sup_{y, \hat{y}_1 \neq \hat{y}_2} \frac{|\ell(y, \hat{y}_1) - \ell(y, \hat{y}_2)|}{|\hat{y}_1 - \hat{y}_2|} < \infty$$

and the uniform convexity condition

$$\inf_{\epsilon > 0} \frac{1}{\epsilon^2} \inf_{y, |\hat{y}_1 - \hat{y}_2| \geq \epsilon} \left(\frac{\ell(y, \hat{y}_1) + \ell(y, \hat{y}_2)}{2} - \ell\left(y, \frac{\hat{y}_1 + \hat{y}_2}{2}\right) \right) > 0.$$

Then there is a constant c *for which the following holds. Suppose that* $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ *is convex and a subset of a* d*-dimensional linear space. Let the distribution of the random pairs* $(X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)$ *be such that the minimizer* $f^* \in \mathcal{F}$ *of* $R(f) = \mathbf{E}(\ell(Y, f(X))$ *exists. Then for all* $x > 0$*, with probability at least* $1 - e^{-x}$*,*

$$\sup_{f \in \mathcal{F}} \left((R(f) - R(f^*)) - 2 \left(\hat{R}(f) - \hat{R}(f^*) \right) \right) \leq c \left(\frac{x + d \log(n/d)}{n} \right),$$
$$\sup_{f \in \mathcal{F}} \left(\left(\hat{R}(f) - \hat{R}(f^*) \right) - 2 \left(R(f) - R(f^*) \right) \right) \leq c \left(\frac{x + d \log(n/d)}{n} \right).$$

Clearly, for the empirical minimizer $\hat{f} = \arg\min_{f \in \mathcal{F}} \hat{R}(f)$, the term $\hat{R}(\hat{f}) - \hat{R}(f^*)$ in the first inequality of the theorem is non-positive, and

so with probability at least $1 - e^{-x}$,

$$R(\hat{f}) \leq \inf_{f \in \mathcal{F}} R(f) + c \left(\frac{x + d \log(n/d)}{n} \right).$$

Results of this kind were first shown for the quadratic loss and function classes with small covering numbers (Lee et al., 1996), later generalized to strictly convex normed spaces and classes with small covering numbers (Mendelson, 2002) and strictly convex losses and classes with small local Rademacher averages (Bartlett et al., 2005, 2006); see also (Koltchinskii, 2006). There are many examples for which these bounds on estimation error decrease at essentially the best possible rate (see, for example, Massart, 2000).

While these results show that the risk of the empirical minimizer converges to the minimal risk in the function class surprisingly quickly, it is not clear that the results are useful for complexity penalization. Is it possible to perform model selection by choosing the class with the smallest penalized empirical risk, and will this lead to an oracle inequality with the corresponding fast rate? It has been suggested that this is not possible: see, for example, the comments in Section 2.1 of (Blanchard et al., 2003). Indeed, Theorem 2 shows that, although the empirical minimizer's risk approaches the minimal risk roughly as n^{-1} in the example of Theorem 3, we cannot hope to *estimate* the minimal risk (or the risk of any function in the class with non-constant loss) to better accuracy than $n^{-1/2}$.

In this note, we give a simple proof that, for nested hierarchies, the fast rates in inequalities of the kind that appear in Theorem 3 *do* imply oracle inequalities for complexity penalization methods. Thus, although the empirical risks that are used to compare classes in the hierarchy fluctuate on a scale that can be large compared to the complexity penalties, for nested hierarchies the fluctuations are sufficiently correlated that they do not excessively affect our choice of class.

2 Oracle inequalities

As above, consider the following complexity penalization approach. Decompose a class $\mathcal{F}$ of real-valued functions defined on an input space $\mathcal{X}$ into a sequence of subsets $\mathcal{F}_1, \mathcal{F}_2, \ldots$. For each j, define the empirical risk minimizer and the true risk minimizer in the class $\mathcal{F}_j$ as

$$\hat{f}_j = \arg\min_{f \in \mathcal{F}_j} \hat{R}(f),$$

$$f_j^* = \arg\min_{f \in \mathcal{F}_j} R(f).$$

Assign to each class $\mathcal{F}_j$ a complexity penalty p_j. Then choose $f_n = \hat{f}_{\hat{m}}$, where

$$\hat{m} = \arg\min_j \left(\hat{R}(\hat{f}_j) + p_j \right).$$

Theorem 4. *Suppose that the classes are ordered by inclusion, choose positive*

numbers ϵ_k that are similarly ordered,

$$\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \mathcal{F}_3 \subseteq \cdots$$

$$\epsilon_1 \leq \epsilon_2 \leq \epsilon_3 \leq \cdots$$

and choose a penalty $p_k = 7\epsilon_k/2$. Then if the data is such that

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(R(f) - R(f_k^*) - 2\left(\hat{R}(f) - \hat{R}(f_k^*)\right) - \epsilon_k \right) \leq 0, \tag{1}$$

$$\sup_k \sup_{f \in \mathcal{F}_k} \left(\hat{R}(f) - \hat{R}(f_k^*) - 2\left(R(f) - R(f_k^*)\right) - \epsilon_k \right) \leq 0, \tag{2}$$

this implies that

$$R(f_n) \leq \inf_k \left(R(f_k^*) + 9\epsilon_k \right).$$

Notice as before that the inequalities (1) and (2) involve random variables. Thus, when they hold with high probability (such as under the conditions of Theorem 3), then the final inequality of the theorem holds with high probability.

If we apply (1) to the empirical minimizer $\hat{f}_k$ in a single class $\mathcal{F}_k$, the difference $\hat{R}(\hat{f}_k) - \hat{R}(f_k^*)$ is non-positive, so we obtain

$$R(\hat{f}_k) \leq \inf_{f \in \mathcal{F}_k} R(f) + \epsilon_k.$$

The theorem shows that adding an $O(\epsilon_k)$ penalty to the empirical risks allows us to choose the class k that satisfies the best of these inequalities. The intuition behind condition (1) is that for functions with small regret

(that is, the difference between their risk and the minimal risk over the class is small), the empirical regret should be a more accurate upper bound on the regret, whereas it need not be so accurate for functions with large regret. Condition (2) provides corresponding lower bounds. This second condition is necessary to ensure that the penalized estimate does not choose the class $\hat{m}$ too large.

3 Proof

We set $p_j = c\epsilon_j$ for some constant c, and we shall see later that $c = 7/2$ suffices.

First we show that $f_n = \hat{f}_{\hat{m}} \in \mathcal{F}_{\hat{m}}$ will have risk that is not much worse than that of any $\hat{f}_j$ from a larger class, $\mathcal{F}_j \supseteq \mathcal{F}_{\hat{m}}$.

Lemma 5. *In the event that $j \geq \hat{m}$,*

$$R(f_n) \leq R(f_j^*) + \max\{2c + 2, 3\}\, \epsilon_j.$$

Proof. From the condition (1), we see that for any j,

$$\hat{R}(f_j^*) \leq \hat{R}(\hat{f}_j) + \frac{\epsilon_j}{2}. \tag{3}$$

In addition, the definition of $f_n = \hat{f}_{\hat{m}}$ implies

$$\hat{R}(\hat{f}_{\hat{m}}) + p_{\hat{m}} \leq \hat{R}(\hat{f}_j) + p_j.$$

Thus, we have

$$\begin{aligned}
\hat{R}(f^*_{\hat{m}}) &\leq \hat{R}(\hat{f}_{\hat{m}}) + \frac{\epsilon_{\hat{m}}}{2} \\
&\leq \hat{R}(\hat{f}_j) + p_j - p_{\hat{m}} + \frac{\epsilon_{\hat{m}}}{2} \\
&\leq \hat{R}(\hat{f}_j) + c\epsilon_j + \max\left\{\frac{1}{2} - c, 0\right\}\epsilon_{\hat{m}} \\
&\leq \hat{R}(\hat{f}_j) + \max\left\{c, \frac{1}{2}\right\}\epsilon_j,
\end{aligned}$$

since $\epsilon_{\hat{m}} \leq \epsilon_j$. Applying (1) for the class $\mathcal{F}_j$, we have

$$\begin{aligned}
R(f^*_{\hat{m}}) &\leq R(f^*_j) + 2\left(\hat{R}(f^*_{\hat{m}}) - \hat{R}(f^*_j)\right) + \epsilon_j \\
&\leq R(f^*_j) + 2\left(\hat{R}(\hat{f}_j) - \hat{R}(f^*_j)\right) + \max\{2c+1, 2\}\,\epsilon_j \\
&\leq R(f^*_j) + \max\{2c+1, 2\}\,\epsilon_j,
\end{aligned}$$

by definition of $\hat{f}_j$. Thus, applying (1) again, this time for the class $\mathcal{F}_{\hat{m}}$, we have

$$\begin{aligned}
R(\hat{f}_{\hat{m}}) &\leq R(f^*_{\hat{m}}) + 2\left(\hat{R}(\hat{f}_{\hat{m}}) - \hat{R}(f^*_{\hat{m}})\right) + \epsilon_{\hat{m}} \\
&\leq R(f^*_j) + \max\{2c+2, 3\}\,\epsilon_j.
\end{aligned}$$

□

The next step is to show that the risk of $f_n = \hat{f}_{\hat{m}}$ cannot be much worse than that of any $\hat{f}_j$ from a smaller class $\mathcal{F}_j \subseteq \mathcal{F}_{\hat{m}}$. First, we show that the risk of the optimal function from $\mathcal{F}_{\hat{m}}$ is not much bigger than the risk of the optimal function from any smaller class.

Lemma 6. *In the event that $j \leq \hat{m}$,*

$$R(f^*_{\hat{m}}) \leq R(f^*_j) + \left(\frac{3}{4} - \frac{c}{2}\right)\epsilon_{\hat{m}} + \frac{c\epsilon_j}{2}.$$

Although it is immediate that $R(f^*_{\hat{m}}) \leq R(f^*_j)$ in this case, we shall see that we need to ensure that, for a suitably large penalty, the gap decreases with $\epsilon_{\hat{m}}$. This is where the condition (2) is important: it allows us to show that $\hat{m}$ is not chosen too large.

Proof. We have

$$\begin{aligned}
&R(f^*_j) - R(f^*_{\hat{m}}) \\
&\geq \frac{1}{2}\left(\hat{R}(f^*_j) - \hat{R}(f^*_{\hat{m}})\right) - \frac{\epsilon_{\hat{m}}}{2} \qquad \text{(by (2), for } \mathcal{F}_{\hat{m}}\text{)} \\
&\geq \frac{1}{2}\left(\hat{R}(\hat{f}_j) - \hat{R}(f^*_{\hat{m}})\right) - \frac{\epsilon_{\hat{m}}}{2} \qquad \text{(by definition of } \hat{f}_j\text{)} \\
&\geq \frac{1}{2}\left(\hat{R}(\hat{f}_{\hat{m}}) + p_{\hat{m}} - p_j - \hat{R}(f^*_{\hat{m}})\right) - \frac{\epsilon_{\hat{m}}}{2} \qquad \text{(by definition of } \hat{m}\text{)} \\
&\geq \frac{1}{2}\left(\hat{R}(f^*_{\hat{m}}) - \frac{\epsilon_{\hat{m}}}{2} + p_{\hat{m}} - p_j - \hat{R}(f^*_{\hat{m}})\right) - \frac{\epsilon_{\hat{m}}}{2} \qquad \text{(by (3))} \\
&= \left(\frac{c}{2} - \frac{3}{4}\right)\epsilon_{\hat{m}} - \frac{c\epsilon_j}{2}.
\end{aligned}$$

The result follows. □

Next we show that this implies the risk of $f_n = \hat{f}_{\hat{m}}$ is not much larger than the risk of any $\hat{f}_j$ from a smaller class.

Lemma 7. *In the event that $j \leq \hat{m}$,*

$$R(f_n) \leq R(f^*_j) + \left(\frac{7}{4} - \frac{c}{2}\right)\epsilon_{\hat{m}} + \frac{c\epsilon_j}{2}.$$

Proof. From (1) and the previous lemma,

$$\begin{aligned} R(f_n) = R(\hat{f}_{\hat{m}}) &\leq R(f^*_{\hat{m}}) + 2\left(\hat{R}(\hat{f}_{\hat{m}}) - \hat{R}(f^*_{\hat{m}})\right) + \epsilon_{\hat{m}} \\ &\leq R(f^*_{\hat{m}}) + \epsilon_{\hat{m}} \\ &= R(f^*_j) + \left(\frac{7}{4} - \frac{c}{2}\right)\epsilon_{\hat{m}} + \frac{c\epsilon_j}{2}. \end{aligned}$$

□

Choosing $c = 7/2$ gives the result.

Acknowledgements

We gratefully acknowledge the support of the NSF under award DMS-0434383. Thanks also to three anonymous reviewers for useful comments that improved the presentation.

References

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

Bartlett, P. L., Bousquet, O., and Mendelson, S. (2005). Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537.

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

Bartlett, P. L. and Mendelson, S. (2006). Empirical minimization. *Probability Theory and Related Fields*, 135(3):311–334.

Blanchard, G., Lugosi, G., and Vayatis, N. (2003). On the rate of convergence of regularized boosting classifiers. *Journal of Machine Learning Research*, 4:861–894.

Dudley, R. M. (1999). *Uniform Central Limit Theorems*. Cambridge University Press.

Koltchinskii, V. (2006). Local Rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656.

Lee, W. S., Bartlett, P. L., and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132.

Massart, P. (2000). Some applications of concentration inequalities to statistics. *Annales de la Faculté des Sciences de Toulouse*, IX:245–303.

Mendelson, S. (2002). Improving the sample complexity using global data. *IEEE Transactions on Information Theory*, 48(7):1977–1991.