Model selection using Rademacher Penalization

Fernando Lozano
Department of Electrical and Computer Engineering
University of New Mexico
Albuquerque,NM 87131 USA
flozano@eece.unm.edu

Abstract

In this paper we describe the use of Rademacher penalization for model selection. As in Vapnik's Guaranteed Risk Minimization (GRM), Rademacher penalization attemps to balance the complexity of the model with its fit to the data by minimizing the sum of the training error and a penalty term, which is an upper bound on the absolute difference between the training error and the generalization error. However, while the GRM penalty is universal, the computation of the Rademacher penalty is *data driven* which means that it depends on the distribution of the data and hence one can expect better performance for particular instances of learning problems. We present experimental evidence that shows that Rademacher penalization can be used as an effective method of model selection in learning problems. In particular we have shown that for the intervals model selection problem, Rademacher penalization outperforms GRM and cross validation (CV) over a wide range of sample sizes. Our experiments also show that the Rademacher penalty resembles more closely the behavior of the absolute difference between generalization error and training error.

Keywords: model selection, Rademacher penalization, generalization error, complexity regularization, VC dimension.

1 Introduction

In the setting of learning from examples one wishes to infer an unknown functional relation between input and output variables from a finite set of examples. Usually, we want to find the function within a function set, that best approximates a target function. The selection of an appropriate set of functions is crucial: a set consisting of very simple functions will not contain a good approximation to a complex target function, while a set that is too complex may fit the training data well, but perform poorly outside of the training set. This is especially critical when the number of data samples available is not very large and/or is corrupted by noise.

Model selection algorithms attempt to solve this problem by selecting candidate functions from different function sets with varying (increasing) complexity, and using some criteria to select from that pool of candidates the function that will likely have the smallest generalization error. A general class of model selection algorithms is that of penalty-based methods, in which the fit of the model to the training data and the complexity of the model are balanced by minimizing the (possibly weighted) sum of the training error, and a penalty term that grows with the complexity of the function set. This is the case in Vapnik's *Guaranteed Risk Minimization* (GRM) [11] in which the penalty term is a bound on the difference between the generalization error and the empirical error. This bound is a function of $\sqrt{d/m}$ where d is the VC dimension of the function set [11] and m is the size of the training set.

A different philosophy is adopted in the method of *Cross Validation* (CV)[8, 9] in which the sample set is split into a training set and a testing set. The training set is used to select a hypothesis from each function set, and the function with the smallest error on the testing set is selected. It is argued in [3] that cross validation must be the preferred method when no additional information on the particular learning problem is available. One of the reasons for this preference is the fact that the penalty added to the empirical risk in methods such as GRM is universal, in the sense that it does not depend on the particular distribution of the data or the target function. On the other hand, CV is sensitive to the distribution and the target function, and has a better tracking ability of the generalization error over different learning problems.

In this paper we present some experimental comparisons of the method of Rademacher penalization for model selection with GRM and CV. Rademacher penalties were introduced by Koltchinskii [5] and share some of the desirable properties of the two algorithms mentioned above. As in GRM, Rademacher penalization computes a bound on the absolute difference between the generalization error and the training error. However, the computation of this bound is *data driven*, which means that it depends on the input distribution, and hence one can expect better performance for particular instances of learning problems.

2 Definitions

In this section we introduce some definitions and notation that will be used throughout the paper. Most of them follow the notation in [3].

We consider the problem of approximating a boolean function $f(x) = I_C(x) : X \to \{0,1\}$ where $(X, \mathcal{A}, \mathbf{P})$ is a probability space (i.e. $\mathcal{A}$ is a σ-algebra of sets in X and $\mathbf{P}$ a probability measure on X),and $C \in \mathcal{A}$. [1]

Assume that we are given a nested sequence of hypothesis classes: $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \mathcal{F}_d \subseteq \cdots$. We are interested in finding a function $h(x)$ within one of the classes $\mathcal{F}_i$ that best approximates our target function $f(x)$, in the sense that it minimizes the generalization error

$$\epsilon(h) = \mathbb{P}\{h(x) \neq f(x)\} = \int_X I_{\{h(x)\neq f(x)\}}(x) d\mathbf{P} \quad (1)$$

Most of the times the distribution of the input space is unknown and we are given only a finite set of labeled examples $S = \{x_i, y_i\}_{i=1}^m$. We assume that the input values $x_i \in X$ are independent and identically distributed (i.i.d) according to $\mathbf{P}$, the labels y_i are assigned by our target function $f(x_i)$ and are possibly corrupted by noise (the label is flipped with some probability $\eta \in [0, 1/2)$).

The usual strategy used to find a function that minimizes (1) is to find the hypothesis within each class that minimizes the training error:

$$\hat{\epsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} I_{\{h(x_i)\neq y_i\}}(x_i) \quad (2)$$

and use some criteria (provided by a model selection algorithm) to select from this pool of functions the one that is more likely to have a smaller generalization error. We denote the selected function by $h_{\tilde{d}}$ where $\mathcal{F}_{\tilde{d}}$ is the hypothesis class to which this function belongs.

3 Rademacher penalization

The idea behind Rademacher penalization (as well as behind GRM) is to find the complexity d that minimizes the quantity $\hat{\epsilon}(h_d) + |\epsilon(h_d) - \hat{\epsilon}(h_d)|$ (where h_d is the hypothesis returned by the training error minimization algorithm for the class $\mathcal{F}_d$). In this way, we would know that the generalization error achieved with the hypothesis h_d is close to the minimum training error within that class, and if in addition, the training error is small, this would guarantee a small generalization error. However, we can not compute $\epsilon(h_d)$ because we do not know the target function $f(x)$ nor the input distribution $\mathbf{P}$. Thus, we settle for a bound on the quantity $|\epsilon(h_d) - \hat{\epsilon}(h_d)|$. It has been shown [11] that for a function class $\mathcal{F}_d$ with VC dimension d, the following inequality holds with high probability:

$$\sup_{h\in\mathcal{F}_d} |\epsilon(h) - \hat{\epsilon}(h))| \leq \sqrt{d \log m/m} \quad (3)$$

This inequality gives us a bound that holds uniformly over the whole function class $\mathcal{F}_d$, in particular it holds for the function h_d. This provides the basis for the GRM model selection algorithm, where the optimal complexity is chosen according to the rule:

$$\tilde{d} = \arg\min_d \left\{ \hat{\epsilon}(d) + \frac{d(\frac{\ln(2m)}{d} + 1)}{m}(1 + \sqrt{(1 + \frac{\hat{\epsilon}(d)m}{d(\frac{\ln(2m)}{d} + 1)}}) \right\} \quad (4)$$

Notice that inequality (3) is universal in the sense that it holds for any input distribution, thus we can expect that a bound on the left hand side of (3) that *does depend* on the input distribution should give us a better approximation to $|\epsilon(h_d) - \hat{\epsilon}(h_d)|$. Such *data-driven* penalty is provided by Rademacher penalization.

Let $\sigma_1, \sigma_2, \ldots, \sigma_m$ be a sequence of i.i.d. Rademacher random variables independent of the data[2] $(x_1, \ldots, x_m)$. The Rademacher penalty of the hypothesis class $\mathcal{F}_d$ is defined as:

$$R_m(\mathcal{F}) = \sup_{h\in\mathcal{F}_d} \left| \frac{1}{m}\sum_{i=1}^{m} \sigma_i I_{\{h(x_i)\neq y_i\}}(x_i) \right| \quad (5)$$

Our model selection algorithm chooses the hypothesis $h_{\tilde{d}}$, according to the rule:

$$\tilde{d} = \arg\min_d \{\hat{\epsilon}(d) + R_m(\mathcal{F}_d)\} \quad (6)$$

We argue next that the Rademacher penalty has a number of desirable properties as a complexity penalization term in a model selection algorithm which uses

[1] $I_C(x)$ denotes the indicator function of C: it is equal to one if $x \in C$, and equal to zero otherwise.

[2] A Rademacher random variable takes values $+1$ and -1 with probability $1/2$ each.

training error minimization as the underlying method to select a hypothesis.

The first tool we need is the following lemma, which is drawn from the theory of empirical processes. Let $\Delta_m(\mathcal{F}_d) = \sup_{h\in\mathcal{F}_d} |\epsilon(h) - \hat{\epsilon}(h))|$, then

Lemma 1 *(Symmetrization inequality)*[3]

$$\mathbb{E}(\Delta_m(\mathcal{F}_d)) \leq 2\mathbb{E}(R_m(\mathcal{F}_d)) \tag{7}$$

Notice that the two expectations in inequality (7) are taken over random variables in different spaces, that is, the expectation on the left-hand side is taken over the input sample set $(x_1, \ldots, x_m)$ while the expectation on the right-hand side is taken over the input samples as well as over the Rademacher random variables $(\sigma_1, \ldots, \sigma_m)$.

Thus, if we are able to estimate the expectation of the Rademacher Penalty, we would have a bound on the expectation of the left hand side of inequality (3). Furthermore, if we prove that the values of the random variables $\Delta_m(\mathcal{F}_d)$ and $R_m(\mathcal{F}_d)$ are concentrated around their expectations, we would have that a particular value of the Rademacher penalty will upper bound the absolute difference between the training error and the generalization error, with high probability. The following two lemmas due to Koltchinskii [5] give bounds on the probability of the values of these random variables being away from their expectations.

Lemma 2 *For all $\varepsilon > 0$,*

$$\mathbb{P}\{\mathbb{E}(R_m(\mathcal{F}_d)) \geq R_m(\mathcal{F}_d) + \varepsilon\} \leq e^{-\varepsilon^2 m/2} \tag{8}$$

and,

$$\mathbb{P}\{R_m(\mathcal{F}_d) \geq \mathbb{E}(R_m(\mathcal{F}_d)) + \varepsilon\} \leq e^{-\varepsilon^2 m/2} \tag{9}$$

Lemma 3 *For all $\varepsilon > 0$,*

$$\mathbb{P}\{\mathbb{E}(\Delta_m(\mathcal{F}_d)) \geq \Delta_m(\mathcal{F}_d) + \varepsilon\} \leq e^{-2\varepsilon^2 m} \tag{10}$$

and,

$$\mathbb{P}\{\Delta_m(\mathcal{F}_d) \geq \mathbb{E}(\Delta_m(\mathcal{F}_d)) + \varepsilon\} \leq e^{-2\varepsilon^2 m} \tag{11}$$

Therefore, with probability at least $1 - 2\delta$, for $m \geq 2/\varepsilon^2 \ln(1/\delta)$, $\Delta_m(\mathcal{F})$ and $R_m(\mathcal{F})$ are within ε of their expectations. The idea is then to use one computation of the Rademacher penalty rather than an estimate of its expectation, as the penalty term in our model selection algorithm.

Notice that (5) can be computed, at least in principle. Furthermore, it is shown in [5] that the computation is equivalent to the minimization of the training error on the relabeled data.

Assume that the supremum in (5) can be achieved in the set $\mathcal{F}$. Then, we need to compute:

$$\max_{h\in\mathcal{F}} \left|\sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i)\right| = \max(\max_{h\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i), \min_{h\in\mathcal{F}} - \sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i)) \tag{12}$$

Define a new set of labels z_i as the flipping of each original label with probability 1/2. Then it is easy to show that:

$$\arg\max_{h\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i) = \arg\min_{h\in\mathcal{F}} \sum_{i=1}^{m} I_{\{-z_i \neq h(x_i)\}}(x_i)$$

and,

$$\arg\min_{h\in\mathcal{F}} \sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i) = \arg\min_{h\in\mathcal{F}} \sum_{i=1}^{m} I_{\{z_i \neq h(x_i)\}}(x_i)$$

Therefore, the computation of the Rademacher penalty involves the following steps:

- Flip the label of each sample with probability 1/2 to get a new set of labels z_i.
- Find the function $h_1 \in \mathcal{F}$ that minimizes the empirical error with respect to the set of labels z_i.
- Find the function $h_2 \in \mathcal{F}$ that minimizes the empirical error with respect to the set of labels $-z_i$.
- Compute $\left|\frac{1}{m}\sum_{i=1}^{m} \sigma_i I_{\{y_i \neq h(x_i)\}}(x_i)\right|$ for $h = h_1, h_2$ and select the maximum of these two values as the Rademacher penalty.

Note that the computational complexity of computing Rademacher penalties is at worse the same as that of the training error minimization algorithm that is used to select the hypothesis from each complexity class.

In summary, the bound on $|\epsilon(h_d) - \hat{\epsilon}(h_d)|$ provided by Rademacher penalization is sensitive to the input distribution. We can think of it also as a measure of the complexity of the hypothesis class: a hypothesis class that is too complex will contain a function that labels correctly most of the relabeled samples resulting in a large value of the Rademacher penalty.

[3] A proof of this lemma can be found in [10].

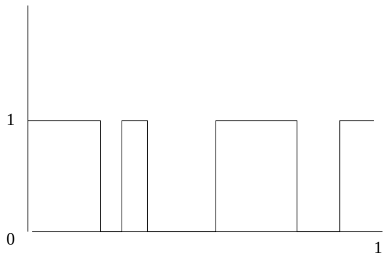

Figure 1: Target function with 6 alternations

4 The intervals model selection problem

In this section we introduce the learning problem used for our experiments. In this problem (referred to as the intervals model selection problem [3]) the input set is the interval $[0,1]$ and the hypothesis class $\mathcal{F}_d$ is the class of binary valued functions over $[0,1]$ with at most d alternations in label. Figure 1 shows an example of a function with 6 alternations.

This is a rare case in which it is possible to find a global minimum of the training error in a reasonable amount of time, as opposed to many learning problems in which finding such minimum is an intractable problem [7, 4, 2, 1]. This makes this problem ideal for analyzing the behavior of our model selection algorithm, since the computation of the Rademacher penalization term requires finding a minimum of the training error with respect to relabeled data and we do not want our results to be obscured by the presence of local minimums . In addition, since we select the target function, it is very easy to compute the generalization error for a given hypothesis, without resorting to other methods such as Monte Carlo integration which can be very time consuming. In fact, the generalization error can be computed exactly for this problem.

We use for our experiments the algorithm described in [3] that yields a running time that is $O(m\log(m))$. This allows us to perform several experiments for different sample sizes, noise rates, and input distributions in a reasonable amount of time[4]. Also, we can compare our results with the results reported previously in [3].

5 Experimental results

In this section we present an experimental comparison of the performance of our model selection algorithm with GRM and cross validation on the intervals model selection problem.

In our experiments we use the target function with 100 intervals corresponding to 99 equally spaced alternations in $[0,1]$. The sample sets were generated from the uniform distribution in $[0,1]$, and corrupted with noise rate of $\eta = 0.2$.

For CV, a the fraction of samples reserved for testing is 10%.

For GRM, the complexity selecting rule is not (4) but rather the following rule which was reported to be more competitive for this problem in [3]:

$$\tilde{d} = \arg\min_d \{\hat{\epsilon}(d) + (d/m)(1 + \sqrt{1 + \hat{\epsilon}(d)m/d})\} \quad (13)$$

We start by showing that at least for this problem, the Rademacher penalty tracks more closely than GRM the behavior of the difference $|\epsilon(d) - \hat{\epsilon}(d)|$. In figures 2 and 3 we plot the value of the penalty term used by each method as a function of complexity for sample sizes of 1000 and 2000 respectively.

We can observe that in general the Rademacher penalty term resembles more closely $|\epsilon(d) - \hat{\epsilon}(d)|$, except for a range of complexities from 0 to approximately 200 in which the GRM curve is closer to the ideal curve. For a smaller sample size both GRM and Rademacher penalization have a very similar behavior in that range. In figures 4 and 5 we plot the penalized errors (training error plus penalty term) for both methods for sample sizes 1000 and 2000. Here again, the Rademacher curve shows a shape much more similar to the generalization error than the GRM curve does. For the larger sample size, the minimum (and hence the complexity selected by the algorithms) is around 100 for both methods. However, for a sample size of 1000, the minimum of the GRM curve has shifted to a smaller value of complexity, while the minimum of the Rademacher curve is still around 100.

Next, we compare the performance of our model selection algorithm to GRM and CV as a function of the sample size. In figure 6 we plot the generalization error of the hypothesis selected by the three methods as a function of the sample size for a range from 50 samples to 2000 samples in steps of 50 samples, averaged over 10 trials of the experiment. We can see that Rademacher penalization outperforms GRM for an initial regime between 50 and 1300, and for larger values of sample size both methods give almost the same generalization error. On the other hand, Rademacher penalization and CV have very similar behavior for sample sizes smaller than 900, and after this value Rademacher penalization outperforms CV. Hence, at least for this problem, Rademacher penalization tracks the other two algorithms in their region of strength, resulting in a better performance over the whole range of sample sizes.

[4] Thanks to Andrew Ng for providing me with a copy of the paper that contains a description of this algorithm

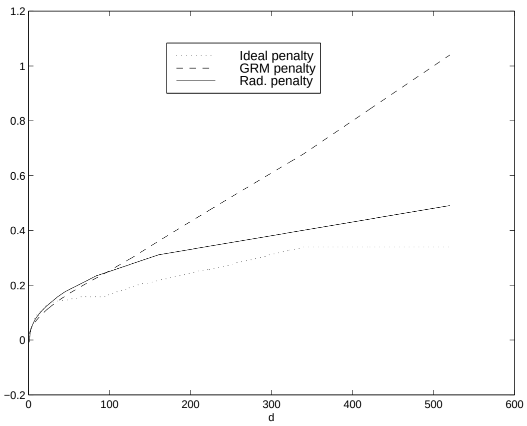

Figure 2: Penalty terms as a function of the complexity d for sample size of 1000. The dotted line shows the "ideal" penalty $|\epsilon(d) - \hat{\epsilon}(d)|$

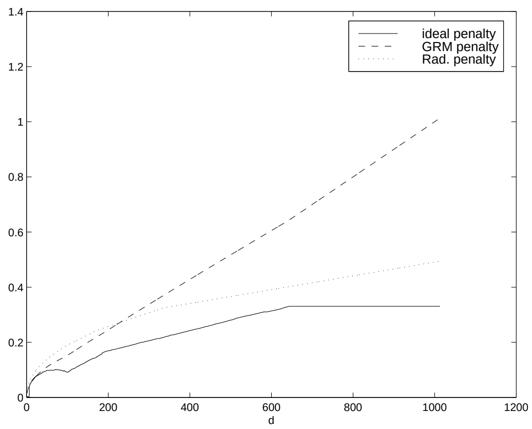

Figure 3: Penalty terms as a function of the complexity d for sample size of 2000.The dotted line shows the "ideal" penalty $|\epsilon(d) - \hat{\epsilon}(d)|$

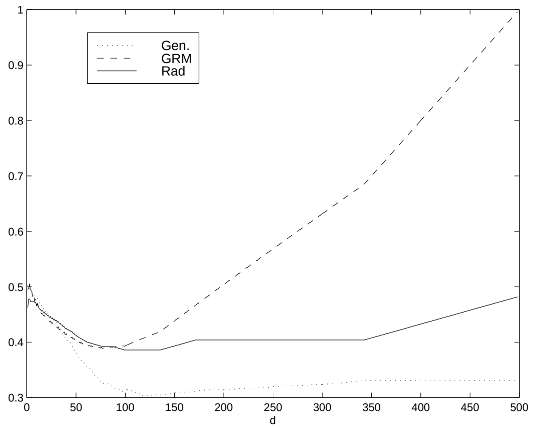

Figure 4: Penalized error as a function of complexity for sample size of 1000. The dotted line shows the actual generalization error of the hypothesis selected from the function class $\mathcal{F}_d$

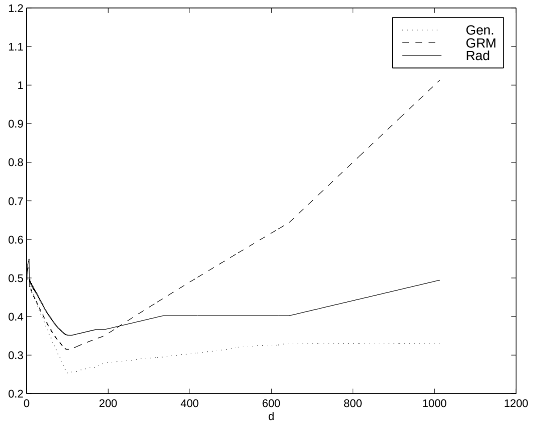

Figure 5: Penalized error as a function of complexity for sample size of 2000. The dotted line shows the actual generalization error of the hypothesis selected from the function class $\mathcal{F}_d$

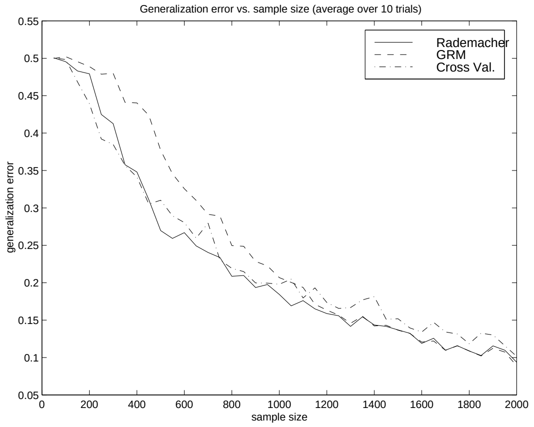

Figure 6: Generalization errors vs. sample size averaged over 10 independent trials.

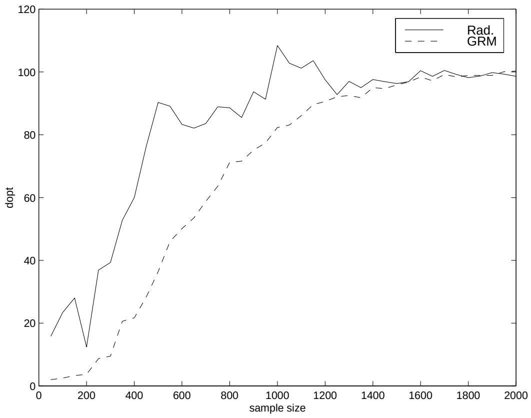

Figure 7: Complexity selected versus sample size, average over 10 trials

It is shown experimentally in [3] that the performance of GRM can not be improved by simple tinkering with the model selection rule. For example by multiplying the penalty term in (13) by a constant less than one, the generalization error in the regime of small to moderate sample sizes goes down, but a prize is paid for larger sample sizes where the generalization error becomes larger. Since Rademacher penalization tracks almost exactly GRM in the regime of larger sample sizes, we can conclude that for this problem Rademacher penalization is preferable to GRM. In [3] the authors show that the poorer performance (with respect to the other algorithms considered there) of GRM is due to the fact that the complexity selected by this method approaches very slowly (as the sample size increases) the "correct" value of d. They wonder if there exists a penalty based algorithm that approaches such value more rapidly than GRM without suffering subsequent overcoding. This is the case of Rademacher penalization, as shown in figure 7 where the value of d selected by both methods is plotted against the sample size.

6 Conclusion

We have presented experimental evidence that demonstrates that Rademacher penalization can be used as an effective method of model selection in learning problems. In particular we have shown that for the intervals model selection problem, Rademacher penalization outperforms GRM over a wide range of sample sizes, and would be preferred to both GRM and CV.

Our experiments also show that the Rademacher penalty resembles more closely the behavior of the absolute difference between generalization error and training error. This fact is important when computing the sample complexity in learning problems. Experiments in applications of learning to control problems have shown that sample complexities based on Rademacher penalization are much smaller than those computed from standard inequalities [6]

References

[1] A. Blum and R. L. Rivest. Training a 3-node neural net is np-complete. In David S. Touretsky, editor, *Advances in Neural Information Processing Systems I*, pages 494–501, San Mateo, CA, 1989. Morgan Kaufmann.

[2] D. R. Hush. Training a sigmoidal node is hard. *Neural Computation*, 11:1249–1260, 1999.

[3] M. Kearns, Y. Mansour, A. Ng, and D. Ron. An experimental and theoretical comparison of model selection methods. *Machine Learning*, 27(1), 1997.

[4] M. Kearns, R. E. Shapire, and L. M. Sellie. Towards efficient agnostic learning. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 341–352, New York, NY, 1992. ACM Press.

[5] V. Koltchinskii. Rademacher penalties and structural risk minimization. *Submitted to IEEE Transactions on Information Theory*, 1999.

[6] V. Koltchinskii, C. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Statistical learning control of uncertain systems: It is better than it seems. Preprint, University of New Mexico, 1999.

[7] L. Pitt and L. Valiant. Computational limitations of learning from examples. *Journal of the ACM*, 35:965–984, 1988.

[8] M. Stone. Cross-validatory choice and assesment of statistical predictions. *Journal of the Royal Statistical Society*, 36, 1974.

[9] M. Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–36, 1977.

[10] A. W. van der Vaart and J. Wellner. *Weak Convergence of Empirical Processes With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

[11] V. Vapnik. *Statistical Learning Theory*. Wiley Interscience, 1998.