# 6.962 Week 4 Summary
# Statistical Learning Theory

Presenter: Emin Martinian

March 1, 2000

## 1 Introduction

Inductive inference is a common principle in a wide variety of disciplines. Broadly speaking, the general inference problem is to understand a relationship between two variables based on empirically observed data. Many inference problems have the form where we are interested in a quantity, $X$, but we are given a noisy observation $Y$. The goal is to build an estimator $\hat{X} = g(Y)$ to minimize some error measure without knowing the density $p(x, y)$.

Statistical Learning Theory (SLT) provides some elegant tools to address both the practical and theoretical aspects of this problem. One of the main tools from SLT is the uniform weak law of large numbers. The uniform weak law characterizes convergence of relative frequencies to the corresponding probabilities. The power and elegance of the uniform weak law is that it is uniform over all distributions. In other words, the uniform weak law can be applied when the probability distribution is unknown provided that another quantity called the VC-dimension can be calculated.

The goal of the talk will be to outline the inductive inference problem. Next we will discuss some pitfalls that commonly occur when solving this problem in practice. We will show how the tools of SLT can be used to understand these pitfalls and develop better inference methods. In the process we will discuss the remarkable uniform weak law of large numbers and the related idea of VC-dimension.

## 2 Deductive And Inductive Inference

If we were given the relationship between $X$ and $Y$ in the form of the joint density $p(x, y)$ then, in principle, we could compute the optimal estimator. The best estimator could be computed by setting up the appropriate constrained optimization and solving analytically or numerically. This is deductive inference in the sense that all the information required to solve the problem is available. Although, the optimal solution might be difficult to calculate, it is computable.

In many practical problems, however, we are not given the appropriate density. Instead we are only given $M$ pairs of data points, $(x_i, y_i)$, drawn according to the unknown i.i.d. density $p(x, y)$. Generally this lack of knowledge makes it difficult to compute an optimal estimator even in principle. A variety of heuristic methods have been proposed and analyzed all of which have the following structure. The parametric set of functions $g(y, \alpha)$ is proposed a priori. The estimator is chosen

by finding the parameter $\alpha^*$ which best matches the data in some sense. The parameter $\alpha$ is an element in the arbitrary set $\Lambda$. Since $\Lambda$ can be an arbitrary set, this form is completely general.

There are various methods of choosing $\alpha$. One of the more straight-forward methods is the Empirical Risk Minimization (ERM) procedure. In the ERM procedure, $\alpha^*$ is the $\alpha \in \Lambda$ such that $g(y, \alpha^*)$ has the smallest estimation error over the training data. A simple example of this procedure is finding the least square fit of a line to the data. In this case the set $\Lambda_L = \mathsf{R}^2$ and the estimator functions are of the form $g(y, \alpha_0, \alpha_1) = \alpha_0 + y \cdot \alpha_1$. The parameters $(\alpha_0, \alpha_1) \in \mathsf{R}^2$ are chosen to minimize the training error $\sum_{i=1}^{M} (g(y_i, \alpha_0, \alpha_1) - x_i)^2$ where $(x_i, y_i)$ are the $M$ training points.

Figure 1 shows an example of finding the best fit line to some randomly generated data. The data was generated according to $Y = \log(1+X)+N$ where $X$ is uniform over $[1, 10]$ and $N$ is zero mean Gaussian with $\sigma = .15$. The $M = 20$ training points used to find the best line fit are represented with open circles. The error of the estimator on the training data is called the training error. The generalization error is the error of the estimator on a set of 2000 additional points selected after the estimator is fixed.

Using the ERM principle for selecting $\alpha^*$ might seem intuitively reasonable. The hope is that if $g(y, \alpha^*)$ has the lowest error on the training set it will also have low error when used to estimate $X$ from $Y$ in later tests. If the set of parameters, $\Lambda$, is well chosen to match the application, the ERM principle can perform quite well. However, we have not provided any mathematical justification for using the ERM principle in the general case where we have no knowledge about $p(x, y)$ other than the training samples.

In fact, there can be serious flaws with the seemingly reasonable ERM principle. To illustrate the flaws consider the least square line fit example discussed previously. The relationship between $X$ and $Y$ may be more complicated than a line so we might want to use a higher order best fit curve such as a parabola. Thus we could broaden our parameter set to $\Lambda_P = \mathsf{R}^3$ where $g(y, \alpha) = \alpha_0 + y \cdot \alpha_1 + y^2 \cdot \alpha_2$. Figure 2 shows an example of finding the best fit parabola.

Since a line is just a special case of a parabola $\Lambda_L \subset \Lambda_P$. Therefore allowing a best fit parabola includes the best fit line as a special case, but also allows more general functions. Consequently the empirical training error for the best estimator from $\Lambda_P$ will often be lower than the empirical training error for the best estimator from $\Lambda_L$. This is the case in our example since the training error in Figure 1 is lower than in Figure 2.

Even though the best fit parabola has a lower training error, we see that its generalization error is worse. This result might seem rather surprising at first since the parabola fits the training data better than the line. The ultimate goal is not to fit the training data, though, but to find a good estimator. We choose the estimator which best fits the training data in the hope that it will work well in general. As we see from the figures, this hope is sometimes in vain.

The effect where a model fits the training data quite well but works poorly in practice is sometimes referred to as over-fitting. Over-fitting can be counterintuitive because one might think that increasing the number of possible estimators will allow us to better explain the relationship between $X$ and $Y$. If a best fit line makes a good estimator one might think a parabolic estimator would be

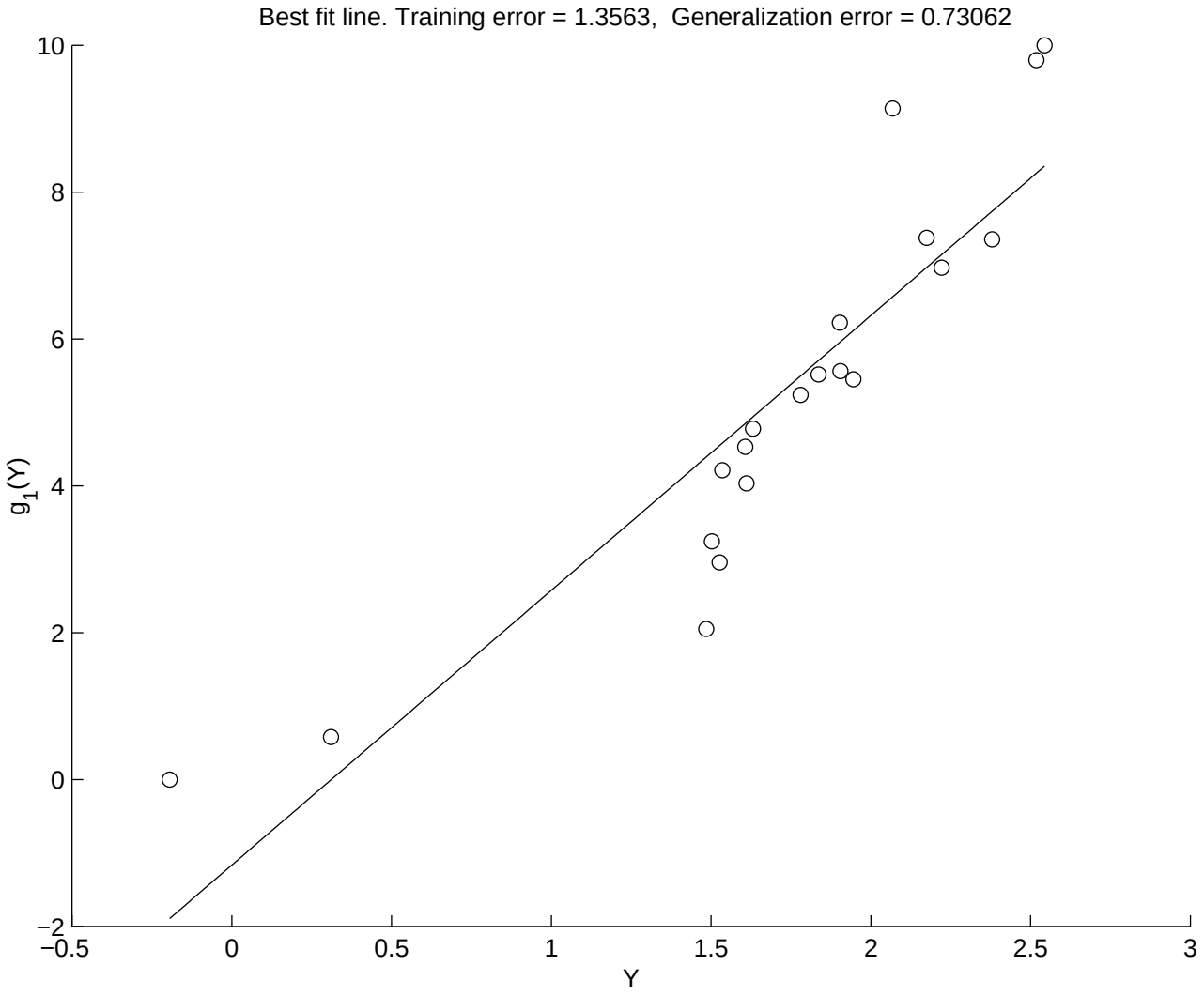


Figure 1: The estimator obtained by doing a best fit line. The data was generated according to $Y = \log(1+X) + N$ where $X$ is uniform over $[1, 10]$ and $N$ is zero mean Gaussian with $\sigma = .15$.

better, and a cubic estimator even better still. Carrying this argument to its conclusion suggests that an $M$th order curve would be the best possible estimator. Indeed, if we let $\Lambda_M = R^M$ so that $g(y, \alpha) = \sum_{i=1}^{M} y^i \cdot \alpha_i$ then we will be able to fit the data *exactly* to make the empirical training error 0. Figure 3 shows an example of finding the best fit $M$th order curve.

Even though the $M$th order curve in Figure 3 has the lowest error on the training data, we can intuitively see that it will not make a good estimator. This is because the $M$th order curve is too complicated a model to infer from the data. The wild gyrations it undergoes to fit the data are unjustified by the small number of training samples available. While the class of $M$th order curves might be able to better explain the relationship between $X$ and $Y$, separating the good estimators from the bad estimators becomes much more difficult. This is an extreme example of the over-fitting effect observed between the best fit line and the best fit parabola.

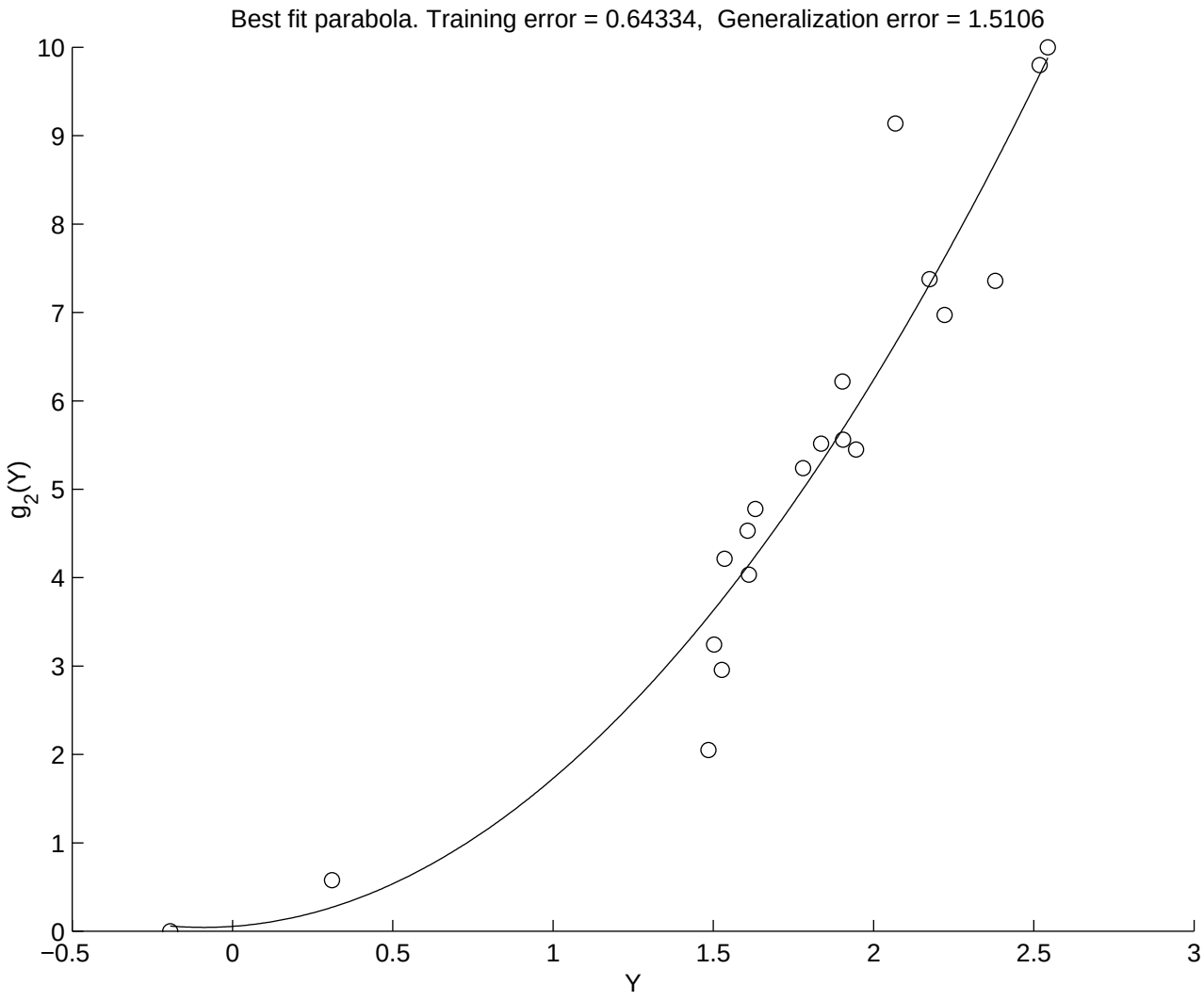


Figure 2: The estimator obtained by doing a best parabola. The data used is the same as in Figure 1.

## 3 Capacity Control

Many other scientists have noticed this over-fitting phenomenon as well. William of Ockham proposed the famous Law of Parsimony as early as the 1300's. The Law of Parsimony also known as Ockham's Razor, essentially says that the simplest explanation is often best. In the previous examples, this seems to be the case. The simplest explanation (the best fit line) has the lowest generalization error even though the relationship between $X$ and $Y$ is nonlinear. On the other hand, we would expect that if significantly more data were available a richer model such as a best fit parabola could perform better.

Ockham's razor is a simple version of the principle of capacity control discussed in [4]. The idea behind capacity control is that one should use a model (i.e. parameter set $\Lambda$) simple enough to be reliably estimated from the data but rich enough to capture as much of the relationship between $X$

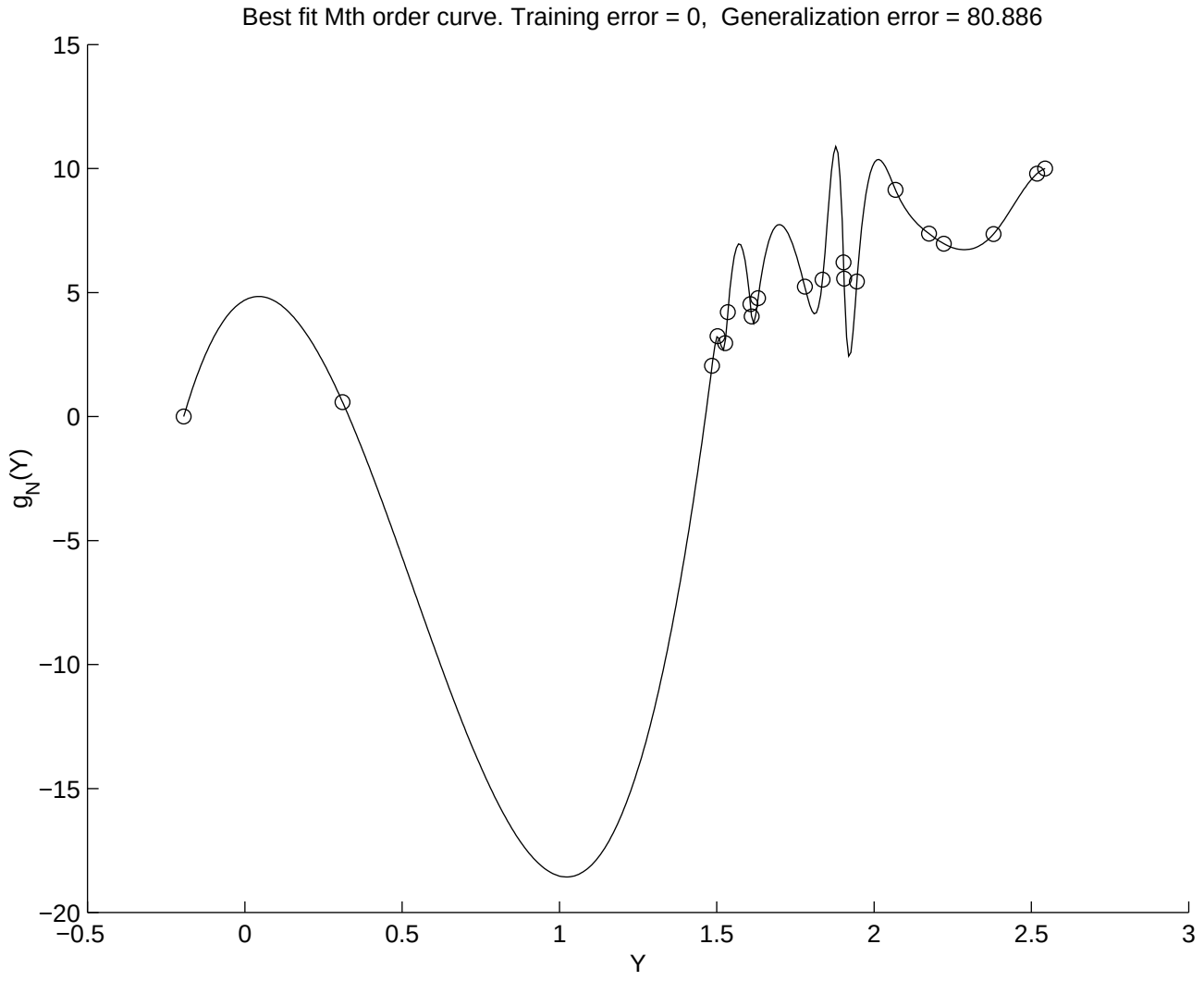


Figure 3: The estimator obtained by doing a best fit $M$th order curve. The data used is the same as in Figure 1 and Figure 2.

and $Y$ as possible.

A popular method of capacity control is the Minimum Description Length (MDL) principle [5]. Essentially, the MDL principle states that we should choose the parameter, $\alpha^*$, to minimize a weighted combination of the training error and the amount of data needed to represent the parameter. This is effectively a more detailed statement of Ockham's razor. It suggests a specific way to trade off the simplicity of the model with its accuracy. In our example, we could represent the description length as the order of the curve $M$. We could then choose the best curve by minimizing a weighting combination of the training error and the model order, $M$.

The MDL principle seems plausible and various results have been found which suggest it can be good asymptotically or in cases where some information about the density $p(x, y)$ is available [5]. For a finite amount of training data, however, these asymptotic results do not apply and the MDL

principle should be considered simply a well motivated heuristic.

## 4 Statistical Learning Theory

The main insight of Statistical Learning Theory is a non-asymptotic characterization of the idea of capacity control. To illustrate this concept we will consider a very simple inductive inference problem. Once again we are given $M$ samples of training data $(x_i, y_i) \in \mathcal{X} \times \mathcal{Y}$. We wish to find the best parameter $\alpha \in \Lambda$. To keep the problem simple, we restrict $\Lambda$, $\mathcal{X}$, and $\mathcal{Y}$ to be a finite sets and we use the probability of error as our classification measure. Specifically, we wish to find the parameter $\alpha^*$ to minimize the probability of classification error. Let $P(\alpha_l)$ be the probability that the estimator $g(\cdot, \alpha_l)$ will be wrong on a random trial. We wish to find $\alpha^*$ to minimize $P(\alpha^*)$.

Unfortunately, we do not have these probabilities available. Let $V(\alpha_l)$ be the fraction of times the estimator $g(\cdot, \alpha_l)$ is wrong on the $M$ training samples. According to the weak law of large numbers, the relative frequency, $V(\alpha_l)$, will converge to the probability $P(\alpha_l)$ asymptotically. Thus if we wait long enough, we can accurately estimate $P(\alpha_l)$ from $V(\alpha_l)$.

This suggests that a reasonable way to choose $\alpha^*$ is to find the estimator which has the minimum training error (i.e. the ERM principle). Assume that the estimator $g(\cdot, \alpha_l)$ has the fewest errors on the training data. The hope behind the ERM principle is that $g(\cdot, \alpha_l)$ will then perform comparably once the training is over. As we saw earlier, the ERM principle can fail badly and the true error of our estimator can be far from the error on the training set.

Consequently, if we choose the estimator $g(\cdot, \alpha_l)$ we are naturally concerned with the probability that $P(\alpha_l)$ will differ by our estimate $V(\alpha_l)$ by more than $\epsilon$. Obviously, if we knew $p(x, y)$ we could calculate this probability exactly. Since $p(x, y)$ is not available, we require some results about convergence of relative frequencies to their probabilities. Using the Chernoff bound we can show

$$\Pr[|V(\alpha_l) - P(\alpha_l)| > \epsilon] \leq 2\exp(-2M\epsilon^2). \tag{1}$$

This bounds the probability that the training error for a particular estimator $g(\cdot, \alpha_l)$ differs from the true error by more than $\epsilon$. This is a useful result, but what we really want to know is the probability that we choose a bad estimator. Essentially we want to bound the probability that the best estimator on the training set, $g(\cdot, \alpha^*)$, has a true error probability, $P(\alpha^*)$, which is much different than the training error $V(\alpha^*)$. Using Equation (1) and the union bound we can show

$$\Pr\left[\sup_{\alpha \in \Lambda} |V(\alpha) - P(\alpha)| > \epsilon\right] \leq 2|\Lambda| \exp[-2M\epsilon^2] = 2\exp\left[-M\left(2\epsilon^2 - \frac{\log|\Lambda|}{M}\right)\right] = 2\exp(-M\eta) \tag{2}$$

where

$$\eta = 2\epsilon^2 - \frac{\log|\Lambda|}{M} \tag{3}$$

and $|\Lambda|$ denotes the cardinality of the finite set $\Lambda$. Note that both Equation (1) and Equation (2) are independent of the distribution $p(x, y)$ and hold for all $M$. Therefore we can bound the probability that we will be badly fooled into choosing $g(\cdot, \alpha_l)$ as our estimator using Equation (2).

The error exponent, $\eta$, in Equation (2) captures the tradeoff between model complexity and successful generalization. If we choose a structure $\Lambda$ with a small number of models we are more likely to choose a good estimator from the set of models. On the other hand, even the best estimator in $\Lambda$ might do poorly at explaining the relationship between $X$ and $Y$.

The Structural Risk Minimization (SRM) procedure discussed in [4] says that we should choose a hierarchy of model classes

$$\Lambda_1 \subset \Lambda_2 \subset ... \subset \Lambda_n \subset ...$$

where $|\Lambda_k| < |\Lambda_{k+1}|$. By choosing an estimator in a small model class we make the probability of choosing a good estimator large, and by choosing a model from a larger model class we allow richer estimators. Thus instead of minimizing the empirical risk which would suggest choosing $g(\cdot, \alpha)$ from a large model class, we minimize combination of the empirical risk and the model complexity. This is similar to the MDL principle, except that instead of a heuristic motivation we now have a rigorous bound to calculate the tradeoff between training error and model complexity.

Thus we see that the model complexity as measured by the number of estimators is an important quantity in inductive inference. Many estimation techniques use countably infinite or even uncountably large parameter spaces, however. For example, the parameter space for the best fit line was $\mathsf{R}^2$. To apply the theory to infinite parameter sets we need to develop a generalization of the result in Equation (2). It turns out that we can effectively replace $|\Lambda|$ with a different measure of the capacity of a set of functions. The correct measure is called the VC-dimension. Specifically, if $V(\alpha)$ denotes the relative frequencies for a set of indicator functions with finite VC-dimension $h$ then

$$\Pr\left\{\sup_{\alpha\in\Lambda}|P(\alpha) - V(\alpha)| > \epsilon\right\} < 4\exp\left\{-M\left[(\epsilon - M^{-1})^2 - \frac{h(1+\log(2M/h))}{M}\right]\right\}. \qquad (4)$$

## References


[1] Vladimir N. Vapnik. An overview of statistical learning theory. *IEEE Transactions On Neural Networks*, 10(5):988–999, September 1999.

[2] Vladimir N. Vapnik. *Estimation Of Dependences Based On Empirical Data.* Springer Series In Statistics. Springer-Verlag, 1982.

[3] Piyush Gupta and P. R. Kumar. The capacity of wireless networks. *IEEE Transactions On Information Theory*, 46(2):388–404, March 2000. Note: This paper contains an interesting use of SLT in the wireless communication setting.

[4] Vladimir N. Vapnik. *Stastical Learning Theory.* John Wiley & Sons, Inc., 1998.

[5] Andrew Barron, Jorma Rissanen, and Bin Yu. The minimum description length principle in coding and modeling. *IEEE Transactions on Information Theory*, 44(6):2743–2760, October 1998.