Learning under persistent drift

Yoav Freund[1] and Yishay mansour[2*]

[1] AT&T Laboratories, 600 Mountain Avenue Murray Hill, NJ 07974-0636 USA
yoav@research.att.com
[2] Dept. of Computer Science Tel Aviv University Tel-Aviv 69978 ISRAEL.
mansour@math.tau.ac.il

Abstract. In this paper we study learning algorithms for environments which are changing over time. Unlike most previous work, we are interested in the case where the changes might be rapid but their "direction" is relatively constant. We model this type of change by assuming that the target distribution is changing continuously at a constant rate from one extreme distribution to another. We show in this case how to use a simple weighting scheme to estimate the error of an hypothesis, and using this estimate, to minimize the error of the prediction.

1 Introduction

One of the oversimplifying assumptions made in the PAC model [Val84] is that all the examples are drawn from the same distribution, and that the target function does not change with time. The drawbacks of this assumption have been widely recognized, and a considerable amount of work was devoted to study the cases where either the distribution [Bar92, BL96] or the target function [HL94, BBDK96] changes over time.

Clearly, without constraints on the way the distribution or target function change over time, it is hopeless to achieve any meaningful learning result. The most common and natural assumption is that the changes are not drastic. A formal way to say this is that the distance between two consecutive distributions (target functions) is bounded by some parameter. This approach was the main subject of previous research [HL94, Bar92, BL96, BBDK96], and has developed interesting learning results.

Common to the results in [HL94, Bar92, BL96] is the assumption that the rate of drift is sufficiently small that the same hypothesis is good for a sufficiently long period of time. Based on this assumption, the learning method of choice is to consider only a certain number of the most recent examples, and use them as though the distribution and target function did not change at all.

The idea of this paper is to show that if we assume that the change in the target distribution is persistent over short periods of time then we can incorporate this knowledge into our algorithm and make good predictions even when

* This research was supported in part by The Israel Science Foundation administered by The Israel Academy of Science and Humanities.

the drift is rapid. We define a model in which the drift in the distribution that governs the generation of examples has a simple structure. We assume that at each time point, we are sampling from a different distribution, but that within a window of size n the changes in the distribution can be approximate well by a linear trajectory. In other words, we denote by $\mathcal{D}_t$ the distribution that governs the generation of examples at time t. This distribution is over both inputs and outputs. Thus it defines both the distribution of inputs and the (probabilistic) relationship between inputs and outputs. Consider some fixed time step t, at which we want to make a prediction. We assume that there is some other distribution $\mathcal{G}_t$ such that the distribution at time $t-i$, for $1 \leq i \leq n$, is approximately equal to $\mathcal{D}_t + i/n(\mathcal{G}_t - \mathcal{D}_t)$. The goal of the learner is to make a good prediction with respect to the distribution $\mathcal{D}_t$, based on the examples that it got from the distributions $\mathcal{D}_{t-n}, \ldots, \mathcal{D}_{t-1}$.

It is not hard to imagine cases in which this assumption will be reasonable. Consider, for example, the problem of predicting the occurance of rain on a particular day as a function of barometric pressure and temperature on the preceding day. As the probability of rain depends on the season, it is reasonable to expect that the best prediction function drifts with time. However, as seasons change relatively slowly, it is also reasonable to assume that the dependence of the expected performance of any fixed function on time can be approximated by a linear function within the range of one month.

One way to compare this new type of assumption to the standard one is to think of them as assumptions about a power series approximation of the way in which probabilities change with time. The standard model corresponds to the assumption about the zero-order or constant term of this expansion. It assumes that the probabilities are changing very slowly. The persistent-drift assumption is an assumption about the first order in the expansion, it assumes that it is the *rate* of change that is more or less constant.

This model has a similar motivation to the model of structured change suggested by Bartlett, Ben David and Kulkarni [BBDK96]. However, we restrict ourselves to the special case in which the change (in both the concept and the input distribution) is a linear function of time. We show that in this case a simple change to the standard method of minimizing the training error yields a simple and effective algorithm.

We design a simple estimator for the error of any fixed hypothesis and give upper bounds on its accuracy. Our algorithm uses this estimator to select a hypothesis with which to make its prediction out of a given hypothesis class. We show that the expected prediction error of this algorithm is never much worse than that of the best hypothesis in the class if the class is finite or has finite pseudo-dimension. We extend this model also to the case where past distributions are only *approximated* by linear drift.

The simplicity of the algorithm and the significance of its performance can be especially appreciated when one considers the fact that it can handle drift in both the input distribution and in the relationship between the input and the output, and that it does not assume that the input-output relationship is perfectly modeled by any of the hypotheses. If the input-output relationship is perfectly modeled by some hypothesis from a known class, and if this dependency does not change with time, then the fact that the input distribution is changing is relatively easy to deal with. Predicting by using the hypothesis that suffered the smallest loss will perform rather well. Intuitively, the worse that can happen is that the old examples might become *irrelevant* because they come from parts of the space that currently have low probability. On the other hand, in our model old examples might be *misleading*, the input-output relationship that approximates them is no longer correct and thus using the hypothesis performs best according to them is a bad idea. As we shall see, our algorithm uses an error estimate that, in effect, uses the old examples as *negative* examples, which cause it to predict in a way that is accurate for the future, rather then for the past.

The paper is organized as follows. In Section 2 we give the precise definition of our model. In Section 3 we derive an unbiased estimator for the error of a hypothesis that has minimal variance. In Section 4.1 we give a bound on the performance of an algorithm that uses a finite set of hypotheses and in Section 4.2 we sketch the derivation of a bound for hypotheses classes of finite pseudo-dimension. We conclude with a summary and some open questions in Section 5.

2 Preliminaries

We assume that the learner is observing a sequence of *examples*. Each example consists of an *input* $x \in X$ and an *output* $y \in [0,1]$,[3] where X is an arbitrary measurable *input space*. We denote the sequence of examples by $(x_0, y_0), (x_1, y_1), \ldots, (x_t, y_t), \ldots$.

Our goal for time t is to predict the output y_t given the input x_t. The examples are assumed to be generated independently at random according to distributions over $X \times [0,1]$. Note that each of these distributions defines both a distribution over the input space and a distribution of outputs for each input $\mathrm{Prob}\{y|x\}$. So far, this setup follows the standard framework of agnostic learning [Vap82, Hau92]. However, the examples are *not* identically distributed, but generated independently at random according to *different* distributions over $X \times [0,1]$. This is the model suggested by Bartlett [Bar92], however, here we make the additional assumption that the distributions are changing with time in a way that can be approximated, for short periods of of time, by a constant rate drift.

More precisely, we denote by $\mathcal{D}_t$ the distribution according to which the t^{th} example, (x_t, y_t) is drawn. We assume that there are parameters $n \in N$ and $1 > \gamma \geq 0$. For any time step $t > n$, there exists a distribution $\mathcal{G}_t$ such that for any $0 \leq i \leq n$:

$$\left\| \mathcal{D}_{t-i} - \left(\mathcal{D}_t + \frac{i}{n}(\mathcal{G}_t - \mathcal{D}_t) \right) \right\|_1 \leq \gamma \;, \tag{1}$$

[3] The output domain can be easily extended to any bounded range of the reals.

where $||\cdot||_1$ denotes the L_1 norm.[4] Intuitively, n corresponds to the time range where linear drift is a good approximation and γ corresponds to the quality of this approximation.

For ease of notation we find it useful to consider a more general setup. We define a function $f:\{0,1,\ldots,n\}\to[0,1]$, such that $f(0)=0$. In the linear case we have $f(i)=i/n$, and rewrite Equation (1) as follows

$$||\mathcal{D}_{t-i}-(f(i)\mathcal{G}_t+(1-f(i))\mathcal{D}_t)||_1\leq\gamma\ . \tag{2}$$

We assume that the learner has access to a hypothesis class $\mathcal{H}$. Each hypothesis $h\in\mathcal{H}$ is a mapping from the input space X to the output space $[0,1]$. The goal of the learner is to predict the t^{th} output ($t>n$) based on the past examples and on the t^{th} input. We compare the performance of the learner with that of the best hypothesis for the t^{th} step.

More precisely, we denote the prediction of the algorithm at time t by $\hat{y}_t$, and the (expected) error of the algorithm at time t by $\epsilon_t^{\text{alg}}=\mathrm{E}\left[|y_t-\hat{y}_t|\right]$, where expectation is taken with respect to the random choice of all examples from time 1 to t. Similarly, for each hypothesis $h\in\mathcal{H}$ we denote by $\epsilon_t(h)$ the error of the hypothesis h at time t. We denote the "minimal achievable error" of the class $\mathcal{H}$ by $\epsilon_t^*=\min_{h\in\mathcal{H}}\epsilon_t(h)$. We measure the performance of our algorithm by the difference $\epsilon_t^{\text{alg}}-\epsilon_t^*$. Our goal is to find algorithms with good guaranteed upper bounds on this performance measure. The performance bounds will hold for each time step t for which the assumptions on the distribution drift hold.

3 Estimating the error of a hypothesis

In this section our goal is to find a good estimate of the error of a fixed hypothesis $h\in\mathcal{H}$ at a fixed time step t when $\gamma=0$. We shall remove the last assumption at the end of the section.

We restrict ourselves to estimators that are linear combinations of past errors. This class of estimators is simple to calculate and to analyze. As we give no lower bounds, the possibility of improved estimators that use other functions of past observations remains open.

Formally, our estimate for the error of a hypothesis h at time t is of the form

$$\hat{\epsilon}_t(h)=\sum_{i=1}^{n}w_i|h(x_{t-i})-y_{t-i}|\ .$$

We call $w_1,\ldots,w_n$ the *weights* and denote by $\mathbf{w}$ the weight vector which consists of these n weights.

[4] In other words

$$||\mathcal{D}_1-\mathcal{D}_2||_1=\int_{X\times[0,1]}|\mathcal{D}_1(x,y)-\mathcal{D}_2(x,y)|d(x,y)\ .$$

First, we find the conditions on $\mathbf{w}$ which should be satisfied in order to make our estimator *unbiased*, i.e. $\epsilon_t(h) = \mathrm{E}\left[\hat{\epsilon}_t(h)\right]$. We denote by $\tilde{\epsilon}_t(h)$ the expected error of the hypothesis h with respect to the distribution $\mathcal{G}_t$. It follows from Equation (2) that

$$\epsilon_{t-i}(h) = f(i)\tilde{\epsilon}_t(h) + (1 - f(i))\epsilon_t(h) \ , \tag{3}$$

for $1 \leq i \leq n$. Thus our requirement is that

$$\begin{aligned} \epsilon_t(h) &= \sum_{i=1}^{n} w_i \left(f(i)\tilde{\epsilon}_t(h) + (1 - f(i))\epsilon_t(h) \right) \\ &= \tilde{\epsilon}_t(h) \left(\sum_{i=1}^{n} w_i f(i) \right) + \epsilon_t(h) \left(\sum_{i=1}^{n} w_i (1 - f(i)) \right) \ . \end{aligned}$$

As we want the choice of weights, $\mathbf{w}$, to be independent of the unknown values of $\tilde{\epsilon}_t(h)$ and $\epsilon_t(h)$ we get the condition that the factor multiplying $\tilde{\epsilon}_t(h)$ must be equal zero, while the factor multiplying $\epsilon_t(h)$ must be equal one. Using vector notation, these conditions can be given as

$$\mathbf{w} \cdot \mathbf{f} = 0 \text{ and } \mathbf{w} \cdot (\mathbf{1} - \mathbf{f}) = 1 \tag{4}$$

where $\mathbf{f} = \langle f(1), \ldots, f(n) \rangle$ and $\mathbf{1} = \langle 1, 1, \ldots, 1 \rangle$. The condition can also be written as

$$\mathbf{w} \cdot \mathbf{1} = 1 \text{ and } \mathbf{w} \cdot \mathbf{f} = 0 \tag{5}$$

Unless $n = 2$, these two conditions leave a lot of freedom in the choice of $\mathbf{w}$. We use this freedom so as to minimize the variance of our estimator. (Actually, we will minimize an upper bound on the variance.) As the estimator is a sum of n independent random variables, its variance is simply

$$\mathrm{Var}\left[\sum_{i=1}^{n} w_i |h(x_{t-i}) - y_{t-i}| \right] = \sum_{i=1}^{n} w_i^2 \, \mathrm{Var}\left[|h(x_{t-i}) - y_{t-i}| \right] \leq \frac{1}{4} \sum_{i=1}^{n} w_i^2 \ .$$

Thus minimizing the L^2 norm of $\mathbf{w}$ is equivalent to minimizing an upper bound on the variance. Minimizing the length of $\mathbf{w}$ under the constraints given in Equation (5) implies that $\mathbf{w}$ is in the span of $\mathbf{1}$ and $\mathbf{f}$,[5] i.e., is of the form

$$\mathbf{w} = a\mathbf{1} + b\mathbf{f} \ .$$

Solving for a and b to satisfy the constraints we find that the choice of $\mathbf{w}$ that gives the unbiased estimator with least variance is

$$\mathbf{w} = \frac{F_2 \mathbf{1} - F_1 \mathbf{f}}{F_2 n - F_1^2}$$

[5] To see why this is so, note that any solution to the conditions in Equation (5) can be written as a sum of the two orthogonal vectors, $\mathbf{w} = \mathbf{u} + \mathbf{v}$, such that $\mathbf{u} = a\mathbf{1} + b\mathbf{f}$ and $\mathbf{v} \cdot \mathbf{1} = \mathbf{v} \cdot \mathbf{f} = 0$. It follows that $\mathbf{v} \cdot \mathbf{u} = 0$ and so $\|\mathbf{w}\|_2^2 = \|\mathbf{u}\|_2^2 + \|\mathbf{v}\|_2^2$ thus the length of $\mathbf{w}$ is minimized if $\mathbf{v} = 0$.

where

$$F_1 = \sum_{i=1}^{n} f(i) \text{ and } F_2 = \sum_{i=1}^{n} f^2(i) \ .$$

For the values of $f(i) = i/n$ that we are interested in, for large n, we have that

$$F_1 = \frac{n+1}{2} \approx \frac{n}{2}$$

$$F_2 = \frac{1}{n^2}\left(\frac{(n+1)^3}{3} - \frac{(n-1)^2}{2} + \frac{n}{6} + \frac{1}{6}\right) \approx \frac{n}{3}$$

To simplify our analysis we use the slightly sub-optimal choice of weights:

$$w_i \doteq 4/n - 6i/n^2 \ .$$

This is not the first time in which a weighted average of past errors has been recommended as a measure of the quality of a hypothesis. Helmbold and Long [HL94] suggested using a similar weighted average to optimize the bounds on the performance of their algorithm. The main difference is in the assumption about the rate of change. While in [HL94] the rate of change is slow, and a single function is suited to fit all the recent examples, we allow rapid changes, and it may be that a function that fits well at the start of the window will fit poorly at its end. This main structural difference manifests itself in the weights. In [HL94] all the weights are positive, and this it corresponds to the assumption that the achievable prediction error is of the same order as the rate of drift. On the other hand, our estimator, which is designed to handle rapid drifts, has some of the weights *negative* (those that correspond to steps $t-n, \ldots, t-(2/3)n$). These negative weights satisfy the need to predict the rate of change of an hypothesis and to fit it well with a linear model. (In order to gain some more intuition, consider two hypothesis that have the same average error on the distributions (each one given the same weight). Since their error changes linearly from one distribution to the other, we would like to pick the hypothesis that has the *highest* error initially, since it will have the lowest error on the last distribution, the one that we are interested in.)

We conclude this section by giving a bound on the probability that our estimator suffers a large error. To do that we use Bernstein's Inequality (see e.g. [Pol84]):

Lemma 1 Bernstein. *Let $Y_1, \ldots, Y_n$ be independent random variables with zero means and and bounded ranges: $|Y_i| \leq M$. Write σ_i^2 for the variance of Y_i. Suppose $V \geq \sigma_1^2 + \cdots + \sigma_n^2$. Then for each $\eta > 0$,*

$$\text{Prob}\left\{|Y_1 + \cdots + Y_n| > \eta\right\} \leq 2\exp\left[-\frac{1}{2}\eta^2/(V + \frac{1}{3}M\eta)\right] \ .$$

In our setting we set $Y_i = w_i|h(x_{t-i}) - y_{t-i}| - w_i\epsilon_{t-i}(h)$, $M = \max w_i$ and $V = 1/4\sum_{i=0}^{n-1} w_i^2$. In such a case $Y_1 + \cdots + Y_n = \hat{\epsilon}_t(h) - \epsilon_t(h)$.

If $M\eta$ is much smaller than V then this bound justifies minimizing V, i.e., the variance of the estimator. For the special case of $f(i) = i/n$, we have $V = 1/n$ and $M = 4/n$, and therefore, for $\eta < 3/4$, we have

$$\text{Prob}\left\{|\hat{\epsilon}_t(h) - \epsilon_t(h)| > \eta\right\} \leq 2e^{-\eta^2 n/4} .$$

Going back to the case that $\gamma > 0$ we derive the following theorem.

Theorem 2. *For the case that $f(i) = i/n$, using $w_i = 4/n - 6i/n^2$, we guarantee that, for any hypothesis h and for any time step t for which the assumptions of Equation (1) hold*

$$\text{Prob}\left\{|\hat{\epsilon}_t(h) - \epsilon_t(h)| > \eta + 2\gamma\right\} \leq 2e^{-\eta^2 n/4} .$$

Proof. First we need to modify equation 3, to state that,

$$|\epsilon_i(h) - f(t-i)\tilde{\epsilon}_t(h) + (1 - f(t-i))\epsilon_t(h)| \leq \gamma .$$

Then we have that,

$$\epsilon_t(h) \leq \tilde{\epsilon}_t(h)\left(\sum_{i=1}^{n} w_i f(i)\right) + \epsilon_t(h)\left(\sum_{i=1}^{n} w_i(1 - f(i))\right) + \gamma\|\mathbf{w}\|_1 ,$$

The theorem follows from an application of Lemma 1, to the case where $f(i) = i/n$, and from the fact that $\sum_{i=0}^{n-1}|w_i| < 2$, for the above choice of weights.

It is interesting to compare this bound to the bound on the deviation of the estimation of the error for the case of i.i.d. drawn examples. In that case, if one uses Hoeffding's bound, one gets a similar bound, but the exponent is $-2\eta^2 n$ instead of $-\eta^2 n/4$. In other words, the sample size that we require, in order to guarantee a given accuracy and reliability level under the linear drift model is only eight times larger than the one required in the analysis standard model of fixed distribution. Most importantly, this bound is independent of the rate of the drift!

4 The learning algorithm

Based on the previous section it is rather straightforward to derive the learning algorithm. Given a sample (x_i, y_i), for $i < t$, we use the last n samples for the algorithm, i.e., (x_i, y_i), for $t - n \leq i \leq t - 1$. For each hypothesis $h \in \mathcal{H}$ we compute $\hat{\epsilon}_t(h)$. The algorithm selects the hypothesis $\hat{h}_t$ that minimizes $\hat{\epsilon}_t(h)$, and returns as its prediction $\hat{h}_t(x_n)$.

In the next subsection we would like to analyze how well does this learning rule does.

4.1 Comparing hypotheses from a finite class

If the size of the class of hypotheses is finite, then a bound on the expected error of the hypothesis generated by our algorithm can be derived by trivial means.

Theorem 3. *At every time step t for which the assumptions of our model hold, the expected error of the algorithm is bounded by*

$$\epsilon_t^{alg} \leq \epsilon_t^* + 2\sqrt{\frac{4}{n}\ln\frac{2|\mathcal{H}|}{\delta}} + 4\gamma + \delta$$

Or, choosing $\delta = c/n$ for some constant $c > 0$, the error is bounded by

$$\epsilon_t^{alg} \leq \epsilon_t^* + c\sqrt{\frac{\ln n|\mathcal{H}|}{n}} + 4\gamma$$

Proof. Assume the opposite. As the algorithm chooses the hypothesis h' whose estimated error is smallest. Denoting the best hypothesis by h^* this means $\hat{\epsilon}_t(h') \leq \hat{\epsilon}_t(h^*)$. This, in turn, implies that either $\hat{\epsilon}_t(h')$ or $\hat{\epsilon}_t(h^*)$ are far from their expected value. From Theorem 2 we know that either of these events can occur with small probability.

4.2 Comparing hypotheses from a class of finite dimension

In the case that the hypothesis class is infinite, we can still bound the error of the hypothesis if we assume that the hypothesis class has finite pseudo-dimension.

In order to prove this we need to go back to the proofs on the uniform convergence of classes of functions and reprove them for this case. We cannot use the existing theorems, because they are stated for the case in which all examples are drawn from the same distribution. Luckily, the proofs given by Haussler [Hau92], which are based on techniques of Pollard [Pol84] do not use the fact that all examples are drawn from the same distribution, and very slight alteration to these proofs lead to the following theorem, which is a slight alteration of Corollary 2 in [Hau92]:

Theorem 4. *Let F be a permissible family of functions from a set Z into [0,1] with pseudo-dimension $d < \infty$. Assume $m \geq 1$. Let $z_1, \ldots, z_n$ be generated independently at random from a sequence of distributions $P_1, \ldots, P_n$, define a $s(f) = \sum_{i=1}^n w_i f(z_i)$ and denote by $\mu(f)$ the expected value of $s(f)$ then for all $0 < \epsilon \leq 1$*

$$\text{Prob}\{\exists f \in F : |s(f) - \mu(f)| > \epsilon\} \leq 8\left(\frac{128e}{\epsilon}\ln\frac{128e}{\epsilon}\right)^d e^{\frac{-\epsilon^2 M}{2304}}$$

where $M = \max_i w_i - \min_i w_i$.

We should comment that no attempt has been made to optimize the constants in this theorem.

Two observations show that the proof given by Haussler [Hau92] can be used essentially verbatim. The proof uses a trick of using two samples and permuting elements among them. The first observation is that all the permutations are done pairwise between elements with the same index in the two samples, thus all the requirements for identical distributions still hold. The second observation is that Hoeffding's bound used in the proof does not require the elements to be identically distributed but only that they are independent.

We use Theorem 4 as follows. We let the set Z be the set of example pairs (x, y) and define the set of functions F to be $|h(x) - y|$ for $h \in \mathcal{H}$. Applying the theorem to our case we get the following theorem:

Theorem 5. *Let $\mathcal{H}$ be a hypothesis class of pseudo-dimension d and assuming that the distribution is drifting with window size n and accuracy $0 \leq \gamma \leq 1$. Then at each time step t the error of the output hypothesis $\hat{h}_t$ is at most,*

$$\epsilon_t^{alg} \leq \epsilon_t^* + c\sqrt{\frac{d}{n}\ln\frac{n}{d}} + 4\gamma \ ,$$

for some constant c.

5 Summary and open problem

We present a new model of learning under changing distributions. In our model the drift is persistent and can be well approximated by a linear drift. We show that a very simple algorithm can achieve good performance even when the drift is rapid.

There are several technical open problems. First, it would be useful to derive some lower bounds on the possible best performance in this model. This would allow us to know how much improvement might be possible. There are many potential places for improving the algorithm. We have restricted our error estimate to be a linear function of past errors and required the estimator to be consistent for the case where $\gamma = 0$. In general, there is no reason to make these restrictions. Moreover, it is clear that our choice of weights is not always optimal. For example, if $\mathcal{G}_t$ is very close to $\mathcal{D}_t$ and γ is relatively large, it is clearly better to assume that the distribution is close to constant and use the *unweighted* average error although it is not consistent.

It is clear that there are many ways in which our analysis can be extended. First, it is clear from our derivation in Section 3 that instead of linear drift one can use many other choices for $f(i)$. Also, it is clear that more general drift structures, which involve convex combinations of more than two distributions can be similarly analyzed. It would be interesting to find choices that correspond to interesting and relevant situations.

References

[Bar92] Bartlett. Learning with slowly changing distribution. In *Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, pages 243–252, 1992.

[BBDK96] Peter Bartlett, Shai Ben-David, and Sanjeev Kulkarni. Learning changing concepts by exploiting the structure of change. In *Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1996.

[BL96] Rakesh D. Barve and Philip M. Long. On the complexity of learning from drifting distributions. In *Proceedings of the Workshop on Computational Learning Theory, Morgan Kaufmann Publishers*, 1996.

[Hau92] David Haussler. Decision theoretic generalization of the pac model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[HL94] David Helmbold and Phill Long. Tracking drifting concepts by minimizing disagreements. *Machine Learning*, 14(1):27–46, 1994. A preliminary version appeared in Proceedings of COLT 1991, 13–23.

[Pol84] David Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

[Val84] Leslie G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, November 1984.

[Vap82] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.