Learning Theory Techniques in Control Theory

Pirkko A. Kuusela

May 20, 1999

Abstract

Learning Theory Techniques in Control Theory

by Pirkko A. Kuusela

Dissertation Director:

Daniel Ocone

We first consider system classification as a learning problem and study a class of mappings from band-limited inputs to outputs realized by continuous-time linear control systems. The difficulty of learning is characterized by the Vapnik-Chervonenkis (VC) dimension of the class of such mappings with binary output classifications. We provide upper and lower bounds for the VC-dimension in terms of the system dimensions amd constants related to controls.

Also, pseudo-dimension bounds are given for studying the input-output behavior without classification. For systems with bounded controls and parameters, fat-shattering bounds are proved in terms of the dimension of the parameter set and the Lipschitz constant associated to the systems.

A control application motivates the problem of learning with side information in which each random sample x gives rise to a translate $s(x)$, where s is a known side information mapping. Both x and $s(x)$ are classified for training, but the aim is to classify correctly only a future unseen x-sample. The learning utilizes non-i.i.d. data and the training and evaluation spaces are different. First we consider a simple problem pointing to phenomena that hold more generally: we calculate exact learning rates for a fixed algorithm for learning an interval on the unit circle under the uniform distribution and reflection as side information.

Typically, learning rates with side information correspond to traditional learning with twice as large a sample, but this may fail for some targets. In general, the advantage of side information depends on the distribution, learning algorithm and due to non-i.i.d. data there is an interaction between the target and the side information mapping.

We incorporate side information in the analysis of uniform convergence of empirical probabilities. Two convergence bounds are analyzed and for measure preserving mappings the exponential improvement in the convergence rate is indicated. As a new technique the bound utilizing Hoeffding's inequality is studied in the large deviations setting. The best improvement doubles the sample size in a bound for consistent algorithms and the sample size is multiplied by 4 in a general convergence bound. However, there are cases in which the exponential improvement due to side information fails.

Contents

Chapter 1

Introduction

1.1 A Short Overview of PAC Learning

We begin with a short overview of Probably Approximately Correct (PAC) learning. This overview is by no means complete and its purpose is to emphasize some aspects that will be the focus of the thesis. For a more detailed discussion we refer to [40], or monographs on the topic, for example [39] and [6].

The basic setting for PAC learning is a set X together with a σ-algebra $\mathcal{S}$ of the subsets of X. We let $\mathcal{P}$ denote the family of probability measures on $(X, \mathcal{S})$. A subset $\mathcal{C} \subseteq \mathcal{S}$ is called a *concept class* and there is a fixed but unknown target concept $T \in \mathcal{C}$ to be learned. The objective is to learn T based on observations that are i.i.d. (independent identically distributed) samples $x_1, \ldots, x_m$ drawn at random from a fixed probability measure $P \in \mathcal{P}$, which may be known or unknown. For each sample an "oracle" will tell us if x_j is in T or not, i.e., the "oracle" returns $\mathbf{1}_T(x_j)$, the value of the characteristic function of T at x_j. After m samples have been drawn the available information is a labeled multisample

$$[(x_1, \mathbf{1}_T(x_1)), \ldots, (x_m, \mathbf{1}_T(x_m))] \in [X \times \{0, 1\}]^m .$$

We form an approximation of the unknown target T by using some algorithm by which we just mean an indexed family of maps $\{A_m\}_{m\geq 1}$ where $A_m : [X \times \{0,1\}]^m \to \mathcal{C}$. Let $h_m = h_m(T;\bar{x})$ be an output of an algorithm with the multisample $[(x_1, \mathbf{1}_T(x_1)), \ldots, (x_m, \mathbf{1}_T(x_m))]$. We measure the goodness of h_m by $d_P(h_m, T) = P(h_m \Delta T)$, where $h_m \Delta T = (h_m \cap T^c) \cup (h_m^c \cap T)$ is the symmetric difference of h_m and T. Observe that $d_P(h_m, T)$ is the probability that h_m will misclassify a future randomly selected sample.

The algorithm $\{A_m\}$ is said to learn the target if $d_P(h_m, T)$ approaches zero as $m \to \infty$. More formally, for PAC learnability we require that

$$r(m,\epsilon) := \sup_{P\in\mathcal{P}} \sup_{T\in\mathcal{C}} P^m\{\bar{x} \in X^m \ ; \ d_P(h_m(T;x),T) > \epsilon\}$$

approaches zero as $m \to \infty$. One can think of $P^m\{\bar{x} \in X^m \ ; \ d_P(h_m(T;x),T) > \epsilon\}$ as a measure of "bad" samples where we call $\bar{x} \in X^m$ "bad" if it gives rise to an estimate with misclassification error larger than ϵ.

Learnability in concept learning is connected to the uniform convergence of empirical probabilities (and in function learning we study the uniform convergence of empirical means). Given a pair $(X, \mathcal{S})$ as above, suppose that $A \subseteq \mathcal{S}$ is a measurable set and our aim is to estimate its probability $P(A)$. By generating i.i.d. samples $x_1, \ldots, x_m \in X$ we let $P_m(A) = 1/m \sum_{i=1}^m \mathbf{1}_A(x_i)$ be the empirical probability of A. For example, Hoeffding's inequality implies that

$$P^m\{\bar{x} \in X^m \ ; \ |P_m(A) - P(A)| > \epsilon\} \leq 2e^{-2\epsilon^2 m}$$

for all P, m and ϵ. Now suppose that a collection of sets $\mathcal{A}$ is given and we wish to estimate simultaneously probabilities for all sets in $\mathcal{A}$ using the same random sample $x_1, \ldots, x_m$. Devroye [17] has shown that

$$P^m\{\bar{x} \in X^m \ ; \ \sup_{A\in\mathcal{A}} |P_m(A) - P(A)| \geq \epsilon\} \leq 4e^{4\epsilon+4\epsilon^2} S(\mathcal{A}, m^2) e^{-2m\epsilon^2},$$

where

$$S(\mathcal{A}, m) = \max_{x_1,\ldots,x_m} \left|\left\{\mathbf{1}_A(x_1), \ldots, \mathbf{1}_A(x_m) \ ; \ A \in \mathcal{A}\right\}\right|$$

is the growth function of the class $\mathcal{A}$. In this bound the exponential term comes from Hoeffding's inequality.

If the empirical probability is known to be zero a better bound can be obtained [18, p. 210] (see [33] for the best bound):

$$P^m\left\{x \in X^m\,;\ \sup_{A\in\mathcal{A};P_m(A)=0} |P_m(A) - P(A)| > \epsilon\right\} \leq 2S(\mathcal{A}, m^2)e^{-m\epsilon},$$

where the exponential term comes from a permutation argument.

Richness of the collection $\mathcal{A}$ can be measured by its Vapnik-Chervonenkis (VC) dimension. A set $S = \{x_1, \ldots, x_n\} \subseteq X$ is said to be *shattered* by $\mathcal{A}$ if, for every subset $B \subseteq S$, there exists a set $A \in \mathcal{A}$ such that $S \cap A = B$. The *Vapnik-Chervonenkis* dimension of $\mathcal{A}$, denoted $\mathrm{VC}(\mathcal{A})$, equals the largest integer n such that there exists a set of cardinality n that is shattered by $\mathcal{A}$. Thus, if $\mathrm{VC}(\mathcal{A}) = d$, $\mathcal{A}$ is not rich enough to distinguish all subsets of any $d+1$ element set, but there is some d element set where subsets can be distinguished. Proving exact values of the VC-dimension is hard and typically one looks for upper and lower bounds for the VC-dimension, as is done in this thesis.

The VC-dimension gives a bound for the growth function $S(\mathcal{A}, m)$. The following lemma is most widely known as Sauer's lemma but it was proved independently in [32], [35] and [38]:

$$S(\mathcal{A}, m) \leq \sum_{i=0}^{\mathrm{VC}(\mathcal{A})} \binom{m}{i} \leq \left(\frac{em}{\mathrm{VC}(\mathcal{A})}\right)^{\mathrm{VC}(\mathcal{A})}.$$

In learning above convergence results are applied to the collection $\mathcal{C}\Delta T$, where $T \in \mathcal{C}$ and also $\mathrm{VC}(\mathcal{C}) = \mathrm{VC}(\mathcal{C}\Delta T)$. Moreover, a concept class $\mathcal{C}$ is distribution-free learnable if and only if it has a finite VC-dimension. We call an algorithm consistent if it agrees with the observed data. If $\mathrm{VC}(\mathcal{C}) = d$ then for any consistent algorithm [33]:

$$r(m, \epsilon) \leq 2\left(\frac{em}{d}\right)^d \epsilon^d e^{2\sqrt{2d}} e^{-m\epsilon}$$

provided that $m \geq 4d/\epsilon$. Given d, ϵ and δ from this we can calculate the *sample complexity*, the required number of samples that guarantees learning with accuracy ϵ and confidence δ.

Similarly, when we wish to learn a fuction $f : X \to [0, 1]$, we can relate the learning problem to the study of uniform convergence of empirical means. In this setting pseudo-dimension and fat-shattering dimension are generalizations of the VC-dimension. In the convergence bounds the role of the growth function $S(\mathcal{A}, m)$ is played by certain covering numbers. Upper bounds for the covering numbers can be obtained in terms of the pseudo-dimension or the fat-shattering dimension and hence they give bounds for the sample complexities. However, those dimensions are not as powerful as the VC-dimension in characterizing learnability.

1.2 Complexity Bounds for Control Systems

In the first part of this thesis we look at system identification as a learning problem. The concept class is a class of mappings from inputs to outputs realized by control systems. We study first the simplest learning setting, the classification problem, and we assume final state sign-observations. That is, given an input, an unknown target system classifies the corresponding final state output as either 0 or 1. The difficulty of learning is characterized by the VC-dimension of the control system concept class.

We study time-invariant linear continuous-time systems $\dot{x} = Ax + Bu$, $y = Cx$, where A, B, C, are $n \times n$, $n \times m$, and $p \times n$ real matrices, and the time interval is [0,1].

Unlike the VC-dimension associated to discrete-time linear systems [26, 14], the VC-dimension for the continuous-time classification problem is unbounded [43]. Therefore we restrict the class of admissible controls and conside controls $u = (u_1, \ldots, u_m)$ such that $u = G\omega$, where the $m \times k$ matrix G parametrizes the control. Control functions $(\omega_1, \ldots, \omega_k)$ are fixed such that each w_i is of the form

$t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t)$ or $t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t)$. Our methods generalize easily to control functions that are linear combinations of the above type of functions.

Control functions $\omega_1, \ldots, \omega_k$ satisfy certain rationality condition associated to the control system (we split the rational function into pieces without poles) and we show that the sign of the final state can by computed by a Boolean formula evaluating polynomial equalities and inequalities. VC-dimension upper bounds for such classes were proved by Goldberg and Jerrum [22]. For scalar sign-observations we prove an upper bound (Theorem 2.1.7)

$$2(2mn^2 + 4n + 1)\log_2\left[8e\left(m2n^2k4(n + \ell_{\max}) + 1\right)\left(6n^2k + 2k + 1\right)\right]$$

where $\ell_{\max}$ is the maximum power of t in $\omega_1, \ldots, \omega_k$. The above bound leads to upper bounds for p-dimensional sign-observations (Theorem 2.1.9) and to the pseudo-dimension upper bounds with loss functions that preserve the rationality structure (see, for example, Theorem 2.1.10).

We prove lower bounds for the VC-dimension with scalar sign-observations. Our interest is the class of mappings from controls (parametrized by G) to sign-observations obtained by varying the system parameters. However, we use duality and study the VC-dimension for mappings from system parameters to sign-observations obtained by varying controls.

By using axis shattering introduced in [14] we prove a lower bound (Theorem 2.1.8):

$$\max\left\{m'\left\lfloor\log_2\left\lfloor\frac{k}{m'}\right\rfloor\right\rfloor, m'\right\}$$

where $m' = \min\{n, k\}$ and $\lfloor x \rfloor$ is the integer part of x. The lower bound is rather general, it holds for all linearly independent control functions and thus no particular form of functions is needed. Observe that as in many applications if the system dimension n is small compared to the "band-width" k, the VC-dimension upper and lower bounds become tighter, both being of the form $C\log_2 k$, with different values of the constant C.

In Section 2.2 we prove an elementary method for obtaining level fat-shattering bounds for function classes that are Lipschitz with respect to parameters, when parameters are in some bounded set of $\mathbb{R}^k$. Assuming that control functions are in L^1 and bounding the system parameters yields a bound for fat-shattering dimension associated to continuous-time linear control systems.

1.3 Learning an Open Interval with Side Information

In the above control setting our interest was the mapping form inputs to sign-observations of the final state. In many cases all intermediate states are also available. As an additional information about the system, one expects that this knowledge should speed the learning of the final state mapping. To accommodate the new problem type we introduce a modification of the learning setting. We assume that each sample gives rise to a new sample $s(x)$, called side information. The "oracle" will classify both x and $s(x)$, and the learning algorithm has access to this new data. However, side information is used only in forming an estimate $\widetilde{h}_m$ of the target T. The goodness of the estimate is evaluated based on its ability to classify correctly a future unseen sample x, but not $s(x)$. This modifies the standard learning setting in two very fundamental ways:

- training utilizes non-i.i.d. data, and
- training and evaluation spaces are different.

As a first step toward understanding the effect of side information, in Chapter 3 we analyze a simple learning problem. This problem has the capacity of allowing exact computations and it also points to phenomena that hold more generally.

We consider an open target interval $T = (a, b) \subset (0, 1)$ and a training scheme in which each x_i drawn from $(0, 1)$ gives rise to $s(x_i) = (x_i + 1/2) \mod 1$. The oracle returns not only $\mathbf{1}_T(x_i)$ but also $\mathbf{1}_T(s(x_i))$ for $i = 1, \ldots, m$. The estimate h_m for the target is the smallest interval containing positive samples.

First we calculate the learning rate

$$\mathrm{Err}(T, h_m, m, \epsilon) = P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\}$$

without side information and compare it to $\mathrm{Err}(T, \widetilde{h}_m, m, \epsilon)$ where $\widetilde{h}_m$ is is an estimate utilizing the original data together with the side information. We show that $\mathrm{Err}(T, h_m, m, \epsilon)$ is independent of T and the density of the probability distribution P (see Theorem 3.2.1 for the learning rate), where as $\mathrm{Err}(T, \widetilde{h}_m, m, \epsilon)$ clearly depends on the underlying probability distribution.

By considering the uniform distribution we establish exact learning rates for each target $T = (a, b)$ with side information $s(x) = (x + 1/2) \mod 1$ (Theorem 3.2.7):

$$\begin{aligned} &P^m\{\bar{x} \in X^m \ ; \ d_P((a, b), \widetilde{h}_m(\bar{x})) > \epsilon\} \\ &\qquad = \begin{cases} (1 - 2\epsilon)^m + 2m\epsilon(1 - 2\epsilon)^{m-1}, & \text{if } \ell \in (\epsilon, \frac{1}{2}] \cup (\frac{1}{2} + \epsilon, 1], \\ 2^m(\ell - \epsilon)^{m-1}\left(\ell - \epsilon + m(1 + \epsilon - 2\ell)\right), & \text{if } \ell \in (\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{2}), \\ 2^m(1 - \ell)^m, & \text{if } \ell \in [\frac{1}{2} + \frac{\epsilon}{2}, \frac{1}{2} + \epsilon], \end{cases} \end{aligned}$$

where $\ell = b - a$.

This result shows a connection between the side information and the target interval—a consequence of learning with non-i.i.d. data. Moreover, in a typical case where $b - a \in (\epsilon, 1/2] \cup (1/2 + \epsilon, 1]$,

$$\frac{\mathrm{Err}(T, \widetilde{h}_m, m, \epsilon)}{\mathrm{Err}(T, h_m, m, \epsilon)} = \frac{P^m\{\bar{x} \in X^m \ ; \ d_P((a, b), \widetilde{h}_m(\bar{x})) > \epsilon\}}{P^m\{\bar{x} \in X^m \ ; \ d_P((a, b), h_m(\bar{x})) > \epsilon\}} \approx \left(\frac{1 - 2\epsilon}{1 - \epsilon}\right)^{m-1}.$$

for m large. As an extreme case when $b - a = 1/2 + \epsilon/2$ we have $\mathrm{Err}(T, \widetilde{h}_m, m, \epsilon)/\mathrm{Err}(T, h_m, m, \epsilon) \approx 1/m$ for m large. That is, in some cases the asymptotic exponential improvement from side information may fail. Note that for $b - a = 1/2 + \epsilon/2$ the effect of non-i.i.d. data is the strongest.

In Section 3.3 we focus on the asymptotic advantage of side information and generalize the probability measures considered as well as the side information mappings. Using large deviation techniques we study the rate at which $\frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}(T, \widetilde{h}_m) \geq \epsilon\}$ converges to zero as $m \to \infty$. By defining a certain set B_ϵ, see (3.3.11), in terms of supports of measures, sample $x_1, \ldots, x_m$ drawn at random from P_g leads to an estimate with misclassification error ϵ or larger if the empirical measure $L_m = 1/m \sum \delta_{x_i}$ is in B_ϵ.

In this setting asymptotic exponential decay rates are computable in terms of entropy functions. In general, upper and lower bounds for the decay are given by $-\inf_{\nu \in \bar{B}_\epsilon} R(\nu \,\|\, P_g)$ and $-\inf_{\nu \in B_\epsilon^0} R(\nu \,\|\, P_g)$, where $R(\nu \,\|\, P_g)$ is the relative entropy of ν with respect to P_g measuring the discrepancy between ν and P_g. Here $\bar{B}_\epsilon$ and B_ϵ^0 denote the closure and the interior of B_ϵ in the weak topology. However, in the examples we consider B_ϵ^0 is empty but by formulating the problem in a stochastic control setting we can prove a better lower bound (Lemma 3.3.2):

$$\liminf_{m \to \infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}(T, \widetilde{h}_m) \geq \epsilon\} \geq -\inf_{\gamma \in B_\epsilon} R(\gamma \,\|\, P_g).$$

The exponential improvement can be characterized as follows (Theorem 3.3.6): Let $\widetilde{h}_m$ be the interior approximation of the target with side information. Assume that samples are from distribution P_g with measurable density function g such that $g(x) > 0$ for Lebesgue almost every x in $[0, 1]$ and the side information mapping $s(x)$ is continuous, one-to-one, onto and $s(I) = I'$ where I and I' are intervals. Then

$$\lim_{m \to \infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}((a, b), \widetilde{h}_m) \geq \epsilon\} = \ln(1 - \epsilon - \epsilon_1),$$

where

$$\epsilon_1 = \inf \left\{ \int g(x) \mathbf{1}_{(s^{-1}(a,c) \cup s^{-1}(d,b)) \setminus ((a,c) \cup (d,b))} dx \;; \quad a < c \le d < b, \int g(x) \mathbf{1}_{(a,c) \cup (d,b)}\, dx = \epsilon \right\}.$$

When $\epsilon_1 > 0$ large deviations show a strict improvement over the learning rate without side information. For example, in the interval problem when $b - a = 1/2 + \epsilon/2$, ϵ_1 will be zero, and the result is in agreement with exact calculations.

Using similar methods we indicate improvement from side information when the learning algorithm is generalized to estimates based on positive and negative samples closest to the endpoints of the target interval. For the precise statement see Theorem 3.3.8.

We finish Chapter 3 with numerical simulations illustrating the small sample effect of side information with different distributions. Overall, we show that given a side information mapping s, the improvement from side information depends on the learning algorithm, distribution, and the target. Thus a worst case analysis will not in general capture a good picture of the side information improvement.

1.4 Side Information in the Learning Analysis

Whereas in Chapter 3 we study a specific learning problem and fixed algorithms, in Chapter 4 our aim is to consider more general learning situations and we discuss how side information can be incorporated into the learning theoretic analysis. Problems arise again from two facts: first the learning utilizes non-i.i.d. data and second the spaces for training and classification are different. We start with the simplest case by assuming that the side information mapping $s(x)$ is measure preserving (this assumption can be relaxed) and this solves the second problem. We work with random variables $\mathbf{1}_A(X) + \mathbf{1}_A(s(X)) \in \{0, 1, 2\}$, where $\mathbf{1}_A$ is the characteristic function of a set A. Depending on the probability of value 2, for example, exponentially faster convergence can be indicated. Note that an assumption on the distribution of $\mathbf{1}_A(X) + \mathbf{1}_A(s(X))$ is a synthesis of all components contributing to the advantage from side information.

In Section 4.1 we study bounds for the uniform convergence of empirical probabilities of the form $C_1 e^{-c_2 m\epsilon^2}$. The exponential term comes from Hoeffding's inequality. As a new technique we relate the exponential term to large deviations via Cramer's theorem in $\mathbb{R}$, see page 102. Unlike large deviations in general, Cramer's theorem in one dimension applies for all m, not only as $m \to \infty$. The exponential convergence rate is given by a convex rate function and one characterization of the rate function is via relative entropies. Next we indicate the best possible exponential improvement. It is obtained if $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 1\}$ for all i or $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{1, 2\}$ for all i (Theorems 4.1.12 and 4.1.18):

$$\Pr\left\{ \sup_{A \in \mathcal{A}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right| \geq \epsilon \right\}$$
$$\leq 4e^{16\epsilon^2 + 16\epsilon} S(\mathcal{A}, 2m^2) e^{-8m\epsilon^2},$$

where $S(\mathcal{A}, m)$ is the growth function of $\mathcal{A}$. In comparison to the result by Devroye [17] the exponential term is improved by a factor of 4 whereas the polynomially growing growth function is worse.

Let p_j denote $\Pr\{\mathbf{1}_A(X) + \mathbf{1}_A(s(X)) = j\}$ for $j = 0, 1, 2$. As A ranges through $\mathcal{A}$ we obtain a corresponding class $\mathbb{P}_{\mathcal{A}} = \{(p_0, p_1, p_2)\}$. If the probability that $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = 2$ is small we show that the improvement from side information does not vanish (see Theorem 4.1.15). If $\mathcal{A}$ is such that $p_2 \leq \beta$ for all $A \in \mathcal{A}$ we can derive the following bound (Theorem 4.1.13):

$$\Pr\left\{ \sup_{A \in \mathcal{A}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right| > \epsilon \right\}$$
$$\leq 4S(\mathcal{A}, 2m^2) \left(\left(1 - C(d, m^2)\right) e^{-m \inf_{p \in \mathbb{P}_{\mathcal{A},d}} \inf_{x \in F_{\epsilon'}} \Lambda_p^*(x)} \right.$$
$$\left. + \, C(d, m^2) e^{-2m\epsilon^2 + 4\epsilon + 4\epsilon^2} \right),$$

where $\mathbb{P}_{\mathcal{A},d} = \{(p_0, p_1, p_2) \ ; \ p_0 + p_1 + p_2 = 1, p_2 \leq \beta + d\}$ and $C(d,m) = 4e^{4d^2+4d} S(\mathcal{A}, 2m^2) e^{-2d^2 m}$.

The term $e^{-m \inf_{p \in \mathbb{P}_{\mathcal{A},d}} \inf_{x \in F_{\epsilon'}} \Lambda^*_{\bar{p}}(x)}$ is the improved rate from Cramer's theorem (see p. 102). $C(d, m^2)$ is a correction term related to the fact that even though p_2 is small, after fixing a sequence of m^2 samples the frequency of 2 is different form p_2. However, it can not be too large from its expected value p_2 and here we apply the uniform convergence result. Thus with high probability $(1 - C(d, m^2))$ the frequency of 2 is less than $p_2 + d$ and the exponential rate is based on this fact. Then in the remaining small probability part we use an exponential term that holds for all values of p_0, p_1, and p_2. We emphasize that the correction term has sample size m^2 and it becomes negligible very quickly.

At the end of Section 4.1.3 we give a numerically verified bound for

$$\inf_{p \in \mathbb{P}_{\mathcal{A},d}} \inf_{x \in F_{\epsilon'}} \Lambda^*_{\bar{p}}(x)$$

together with a numerical comparison on the advantage of side information.

We point out as a separate interest that above methods can be applied in a similar way to the standard learning problems without side information and better exponential terms can be derived if, for example, we know that $P(A) \leq \lambda$ for all $A \in \mathcal{A}$.

If the empirical probability is known to be zero a better bound of the form $C_1 e^{-c_2 m \epsilon}$ can be derived. The exponential term comes from a permutation argument; by incorporating the side information the best bound is when $Z_i^A = \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 1\}$ (Theorem 4.2.2):

$$\Pr\Big\{ \sup_{\substack{A \in \mathcal{A} \\ \sum_{i=1}^m Z_i^A = 0}} |\frac{1}{2m} \sum_{i=1}^m Z_i^A - P(A)| > \epsilon \Big\} \leq 2S(\mathcal{A}, 2m^2) e^{2(1+\epsilon)} e^{-2m\epsilon}.$$

For example, the above situation arises in the interval problem when the target $T = (a, b)$ with $b - a \in (\epsilon, 1/2]$ and the corresponding learning rate is $(1 - 2\epsilon)^m + 2m\epsilon(1 - 2\epsilon)^{m-1}$ showing the tightness of the exponential term in the bound.

The following result indicates improvement when $p_2 \le \gamma < \epsilon$ (Theorem 4.2.3):

$$\Pr\Big\{ \sup_{\substack{A \in \mathcal{A} \\ \sum_{i=1}^m Z_i^A = 0}} |\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)| > \epsilon \Big\}$$
$$\le 2S(\mathcal{A}, 2m^2)\left[\left(1 - C(d, m^2)\right) e^{-mf(m^2\rho)} \right.$$
$$\left. + C(d, m^2) e^{1+\epsilon} e^{-m\epsilon}\right],$$

where

$$f(k) = \frac{k}{m^2} + \frac{2((m\epsilon - 1)(m-1) - k)}{m^2 - k},$$

$\rho = \gamma + d$ and $m^2\rho \le (m\epsilon - 1)(m-1)$.

Again we point to the correction term with sample size m^2. If $\rho = \epsilon$ the above theorem can not indicate an exponential term better than $e^{-m\epsilon}$. However, a situation like this occurs in the interval problem when $b - a = 1/2 + \epsilon/2$ and we showed that the learning rate is $(1-\epsilon)^m$.

We finish the chapter by discussing a possible approach of incorporating side information into learning analysis when original samples and side information samples are fundamentally different objects. The idea is to capture side information as a restrictive condition for consistency. We propose a complexity measure that shares many VC-type properties (for example, Sauer's lemma holds, see Lemma 4.5.2). Also the complexity measure and its data dependent version have a property of "probable smoothness", thus making them fit to the generalized setting of structural minimization by Shawe-Taylor et.al. [34]. However, the restrictive power of side information depends on how "special" the VC-shattered sets are. We illustrate that for hyperplanes, for which shattered sets are by no means "special", adding side information restrictions does not reduce the complexity dimension by more than 1. The reduction power for control systems is an open problem.

Chapter 2

Learning Complexity Bounds for Continuous-Time Linear Control Systems

System identification can be viewed as a learning problem. The aim is to learn a mapping from inputs to outputs given by an unknown control system. This mapping is an element of a concept or function class which consists of all possible mappings induced by control systems having a specified input-output structure. The difficulty of learning in various settings can be characterized by complexity dimensions describing the "richness" of the concept or function class.

In Section 2.1 we study the Vapnik-Chervonenkis (VC) dimension associated to control systems and from that we derive pseudo-dimension bounds. We provide upper and lower bounds for these complexity dimensions in terms of system dimensions and constants related to the control functions. In the last part of the section we formulate the system identification as a learning problem and illustrate how the VC-dimension bounds are applied in the learning setting of structural risk minimization.

In Section 2.2 we focus on fat-shattering, which is a scale sensitive modification of the pseudo-dimension. We show how the fat-shattering dimension can be bounded for Lipschitz functions with parameters in a bounded subset of an Euclidean space. This result is used to derive a fat-shattering bound for control systems. As in the previous section we end by formulating the corresponding learning problem and by providing an estimate for the number of training samples required for given accuracy and confidence.

2.1 VC- and Pseudo-Dimension Bounds

s:vcdim

The learnability of the input-output mapping with sign-observations in the Probably Approximately Correct (PAC) sense is equivalent to the finiteness of the VC-dimension. Pseudo-dimension, a generalization of the VC-dimension, is related to learning real-valued functions. Finiteness of the pseudo-dimension implies learnability in the PAC sense, but the converse is not true, see [39, p. 206] or [8] in general.

In the context of learning we discuss continuous-time linear control systems:

e:system

$$\dot{x} = Ax + Bu, \qquad x(0) = x^0, \qquad y = Cx, \tag{2.2.1}$$

where A, B, and C are $n \times n$, $n \times m$, and $p \times n$ real matrices, and the time interval is [0,1]. We study sign-observations (see [27] for related work in control theory):

$$\text{sign } y(1) = (\text{sign } y_1(1), \ldots, \text{sign } y_p(1))^T ,$$

where sign $z = 0$, if $z \leq 0$, sign $z = 1$, if $z > 0$ and ^{T} stands for the transpose. This is a classification problem; each output is classified either 0 or 1. For scalar observations the VC-dimension can be used to study the complexity of the problem. When $p > 1$, a generalization of the VC-dimension or a loss function is needed.

In general, unlike the VC-dimension associated to discrete-time linear systems [14, 26], the VC-dimension of the classification problem for continuous-time control systems is unbounded [43], even when $n = 1$. Therefore we restrict the class of admissible controls in order to achieve a bound for the VC-dimension. We consider controls $u = (u_1, \ldots, u_m)$ such that

$$u = G\omega,$$

where G is a $m \times k$ matrix that parametrizes the control. The set of control functions $\Omega = \{\omega_1, \ldots, \omega_k\}$ is fixed. The bounds for the VC-dimension or the pseudo-dimension will depend on the properties of the set Ω.

For scalar inputs (i.e., $m = 1$) the VC-dimension associated to the mapping from inputs G to scalar sign-observations is bounded by k, which in fact can be very large in applications. This bound is tight; we give an example of a function class Ω for which the associated VC-dimension is indeed k (see Section 2.1.6). However, by considering band-limited control functions a better bound can be achieved. In this work we consider the following set of control functions

e:contfn

$$\begin{aligned}\Omega = \Big\{\omega_1, \ldots, \omega_k \,;\, &\omega_1, \ldots, \omega_k \text{ linearly independent and} \\ &\omega_j = t^{\ell_j} e^{\alpha_j t} \sin(\beta_j t) \text{ or } \omega_j = t^{\ell_j} e^{\alpha_j t} \cos(\beta_j t) \\ &\text{with } \ell_j \in \mathbb{N}, \alpha_j, \beta_j \in \mathbb{R}, j = 1, \ldots, k\Big\},\end{aligned} \tag{2.2.2}$$

and let

e:ell-max

$$\ell_{\max} = \max\{\ell_1, \ldots, \ell_k\}. \tag{2.2.3}$$

Results stated hold with minor modifications if control functions $\omega_j, j = 1, \ldots, k$ are, for example, linear combinations of the functions of the above form. However, that makes some proofs and formulae a little messier.

We show that the elements of Ω satisfy a certain rationality condition and we study upper and lower bounds for the VC-dimension in terms of the dimensions of the system and constants related to the controls.

For other research combining learning and systems theory see, for example, the articles in the Learning Theory Special Issue of Systems and Control Letters [9]. This issue also contains a compact introduction to some statistical aspects of the PAC learning theory [40]. In addition, CDC'96 had a learning theory session in which Ljung [28] asked if learning theory can help to develop system identification in a parameter free way. Johansen and Wayer [25] have shown steps in that direction. See also the work of Fiechter [21], and the references therein.

In this section, we first present upper bounds for the VC-dimension associated to systems (2.2.1) with scalar observations. After that, lower bounds for systems with a scalar input are developed. We carry further the method used to prove upper bounds to study p-dimensional sign-observations. Finally, we show that VC-dimension bounds can be modified easily to give bounds for the pseudo-dimension, which is related to learning the actual input-output behavior of (2.2.1) (not just the classification problem), and we end by formulating system identification as a learning problem and by providing a bound for the sample complexity.

2.1.1 Definitions and Statements of the Main Results

We recall the definition of the VC-dimension of a concept class $\mathcal{C}$. A subset $D = \{d_1, \ldots, d_n\} \subset X$ is said to be *shattered* by $\mathcal{C}$ if, for every subset $A \subseteq D$, there exists a set $B \in \mathcal{C}$ so that $A = B \cap D$. *The VC-dimension of* $\mathcal{C}$ denoted by $\mathrm{VC}(\mathcal{C})$ is the largest integer n (possibly infinity) such that there exists a set D of cardinality n that is shattered. For example, in $\mathbb{R}^k$ the VC-dimension of closed half-spaces through the origin is k [41].

For our purposes, it is more convenient to work with shattering in terms of dichotomies, i.e., boolean-valued maps. We identify subsets of D with Boolean functions $\phi : D \to \{0, 1\}$. Similarly, each set $C \in \mathcal{C}$ gives rise to a Boolean function on X, and intersections $C \cap D$ are restrictions of functions to D. In this language, a subset $D \subset X$ is shattered by $\mathcal{F} := \{\phi\ ;\ \phi : X \to \{0,1\}\}$, if every dichotomy on D is a restriction to D of some $\phi \in \mathcal{F}$.

Given a class of functions $\mathcal{F} : X \to Y$ and a loss function $L : Y \times Y \to [0, r]$, we introduce for each $f \in \mathcal{F}$ the function

e:pdclass

$$A_{f,L} : X \times Y \times \mathbb{R} \to \{0,1\};\ \ (x, y, \rho) \mapsto \operatorname{sign}(L(f(x), y) - \rho), \tag{2.2.4}$$

and let $A_{\mathcal{F},L}$ denote all such $A_{f,L}$ with $f \in \mathcal{F}$. The *pseudo-dimension of $\mathcal{F}$ with respect to the loss function L*, $\mathrm{PD}[\mathcal{F}, L]$, is defined as:

$$\mathrm{PD}[\mathcal{F}, L] := \mathrm{VC}(A_{\mathcal{F},L}).$$

Techniques for Proving VC-Dimension Bounds

In the context of control theory it is sometimes easier to work with the *dual VC-dimension*. Assume that a function $F : \Lambda \times X \to \{0,1\}$ is given. This induces two function classes

$$\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0,1\}\ ;\ \lambda \in \Lambda\}$$

and

$$\mathcal{F}^* := \{F(\cdot, x) : \Lambda \to \{0,1\}\ ;\ x \in X\}.$$

$\mathrm{VC}(\mathcal{F}^*)$ is called the dual VC-dimension of $\mathcal{F}$ and it is related to $\mathrm{VC}(\mathcal{F})$ as follows [39]:

e:dualvc

$$\mathrm{VC}(\mathcal{F}) \geq \lfloor \log_2 \mathrm{VC}(\mathcal{F}^*) \rfloor, \tag{2.2.5}$$

where $\lfloor x \rfloor$ is the integer part of x.

A sharper estimate can be obtained if Λ can be written as a product $\Lambda_1 \times \cdots \times \Lambda_n$. The following construction and result are due to DasGupta and Sontag [14]. We study in particular those dichotomies that are defined on "rectangular" subsets of

Λ. Let $L = L_1 \times \cdots \times L_n$ be a subset of Λ such that for each i, $L_i \subset \Lambda_i$ is nonempty. Given any index $1 \leq \kappa \leq n$ a *κ-axis dichotomy* on L is any function $\delta : L \to \{0, 1\}$ which depends only on the κth coordinate, i.e., there is some function $\phi : L_\kappa \to \{0, 1\}$ so that $\delta(\lambda_1, \ldots, \lambda_n) = \phi(\lambda_\kappa)$ for all $(\lambda_1, \ldots, \lambda_n) \in L$. We call that a mapping is an axis dichotomy if it is a κ-axis dichotomy for some κ. A rectangular set L is said to be *axis shattered* by $\mathcal{F}^*$ if every axis dichotomy is a restriction to L of some function of the form $F(\cdot, x) : \Lambda \to \{0, 1\}$, for some $x \in X$.

t:axissh

Theorem 2.1.1. (Axis shattering bound [14]) *If $L = L_1 \times \cdots \times L_n \subset \Lambda$ can be axis shattered and each L_i has cardinality $r_i > 0$, then VC($\mathcal{F}$) $\geq \lfloor \log_2(r_1) \rfloor + \cdots + \lfloor \log_2(r_n) \rfloor$.*

Upper bounds for VC-dimensions of concept classes that are obtained by evaluating polynomial equalities and inequalities can be obtained in terms of the number and degrees of the polynomials:

t:gold

Theorem 2.1.2. (Goldberg-Jerrum bound [22]) *Given a function $F : \Lambda \times X \to \{0, 1\}$ and the associated concept class $\mathcal{F} := \{F(\lambda, \cdot) : X \to \{0, 1\} \; ; \; \lambda \in \Lambda\}$, suppose that $\Lambda = \mathbb{R}^\ell$ and $X = \mathbb{R}^k$. Let F be defined in terms of a Boolean formula involving at most s polynomial equalities and inequalities in $\ell + k$ variables, each polynomial being of degree at most d in λ for all $x \in \mathbb{R}^k$. Then, VC($\mathcal{F}$) $\leq 2\ell \log_2(8eds)$.*

The Goldberg-Jerrum bound is based on a result showing that the number of sign assignments $\{-1, 0, 1\}$ to polynomials can not grow too fast:

t:signass

Theorem 2.1.3. *([22]) Suppose that $f_1, \ldots, f_m$ are polynomials of degree at most d in $n \leq m$ variables. Then the number of distinct vectors*

$$\left[\text{sign } f_1(x), \ldots, \text{sign } f_m(x)\right]' \in \{-1, 0, 1\}^m,$$

that can be generated by varying x over $\mathbb{R}^n$, is at most $((8edm)/n)^n$.

Linear Systems

Our main results are based on the fact that the elements of Ω satisfy a certain rationality condition.

Definition 1. (Rationality condition (RAT)) Let n be a positive integer. We say that a bounded function $\omega : [0,1] \to \mathbb{R}$ satisfies the rationality condition relative to the class of n-dimensional systems if there exists h polynomial functions $f_1, \ldots, f_h : \mathbb{R}^4 \to \mathbb{R}$ and $2hn$ rational functions $r_{ij\tilde{\ell}}$, $i \in \{1,2\}$, $j \in \{1, \ldots, h\}$ and $\tilde{\ell} \in \{1, \ldots, n\}$ with no poles on subsets $S_{ij\tilde{\ell}}$ of $\mathbb{R}^4$, such that the following properties hold:

1. For each $i \in \{1,2\}$, $\tilde{\ell} \in \{1, \ldots, n\}$, $\mathbb{R}^4$ is a disjoint union of $S_{i1\tilde{\ell}}, \ldots, S_{ih\tilde{\ell}}$.

2. Each $S_{ij\tilde{\ell}}$ can be defined in terms of a Boolean expression involving $[f_1 = 0]$, $\ldots$, $[f_h = 0]$, where we say that for functions $f_1, \ldots, f_h : \mathbb{R}^4 \to \mathbb{R}$, $[f_i = 0]$ has value 1, if $f_i(x_1, x_2, x_3, x_4) = 0$, and 0 otherwise.

3. Letting $r_{i\tilde{\ell}} : \mathbb{R}^4 \to \mathbb{R}$, $i \in \{1,2\}$, $\tilde{\ell} \in \{1, \ldots, n\}$ be defined as

$$r_{i\tilde{\ell}}(v) = \begin{cases} r_{i1\tilde{\ell}}(v), & \text{if } v \in S_{i1\tilde{\ell}}, \\ \vdots & \vdots \\ r_{ih\tilde{\ell}}(v), & \text{if } v \in S_{ih\tilde{\ell}}, \end{cases}$$

then for each $a, b \in \mathbb{R}$, and for all $\tilde{\ell} \in \{1, \ldots, n\}$

$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \cos(bt) \omega(t) dt = r_{1\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b),$$
$$\int_0^1 t^{\tilde{\ell}-1} e^{at} \sin(bt) \omega(t) dt = r_{2\tilde{\ell}}(a, b, e^a \cos b, e^a \sin b).$$

We denote by $d_{\max}$ the maximum degree of any polynomial (i.e., $f_1, \ldots, f_h$, numerators and denominators of $r_{ij\tilde{\ell}}$'s) appearing in the rationality condition .

Remark 2.1.4. First, entries of e^{At} are functions of the form $t^s e^{at}\cos(bt)$ and $t^s e^{at}\sin(bt)$. Solving (2.2.1) involves convolutions of e^{At} and the input functions ω_j, and we require those to be rational functions.

e:rat-cond
Example 2.1.5. Let $\omega(t) = \sin(ct)$, with nonzero c. Then

$$\int_0^1 e^{at}\sin(bt)\omega(t)dt = \frac{1}{2}\int_0^1 e^{at}\cos((b-c)t)dt - \frac{1}{2}\int_0^1 e^{at}\cos((b+c)t)dt$$

after integration this can be split into cases with no poles yielding

$$\int_0^1 e^{at}\sin(bt)\omega(t)dt = \begin{cases} \frac{p(a,b,e^a\cos b, e^a\sin b)}{(a^2+(b+c)^2)(a^2+(b-c)^2)}, & \text{if } f_1 \neq 0, f_2 \neq 0, \\ \frac{b-\sin b\cos b}{2b}, & \text{if } f_1 = 0, f_2 \neq 0, \\ \frac{\sin b\cos b - b}{2b}, & \text{if } f_1 \neq 0, f_2 = 0, \end{cases}$$

where

$$f_1(a,b,e^a\sin b, e^a\cos b) = a^2 + (b+c)^2,$$
$$f_2(a,b,e^a\sin b, e^a\cos b) = a^2 + (b-c)^2,$$

and $p(a,b,e^a\cos b, e^a\sin b)$ stands for the polynomial

$$\begin{aligned} &-4abc + 4abce^a\cos b\cos c - 2a^2ce^a\cos c\sin b + 2b^2ce^a\cos c\sin b \\ &\quad - 2c^3e^a\cos c\sin b - 2a^2be^a\cos b\sin c - 2b^3e^a\cos b\sin c + 2bc^2e^a\cos b\sin c \\ &\qquad + 2a^3e^a\sin b\sin c + 2ab^2e^a\sin b\sin c + 2ac^2e^a\sin b\sin c. \end{aligned}$$

l:input-rat
Lemma 2.1.6. *Each $\omega_j \in \Omega$ given by (2.2.2) satisfies the rationality condition. Further, h, the number of pieces in (RAT), is at most 3 and $d_{\max}$, the maximum degree of polynomials in (RAT), is at most $4(n+\ell_{\max})$, where $\ell_{\max}$ is given by (2.2.3).*

We recall briefly the control system setting. We study systems

e:learn-syst
$$\begin{aligned} &\dot{x} = Ax + Bu, \quad x(0) = x^0, \quad y = Cx, \\ &u = G\omega, \quad \text{and} \quad \omega_j \in \Omega \text{ for } j = 1,\ldots,k. \end{aligned} \tag{2.2.6}$$

d:csclass

Definition 2. (Sign system concept class, $\mathcal{F}_{m,p}$) Order the set of control functions Ω and denote $\omega = (\omega_1, \ldots, \omega_k)^T$. Let

$$X_\Omega = \{G\omega : [0,1] \to \mathbb{R}^m \;;\; G \in \mathbb{R}^{mk}\}$$

and for each linear system $\Sigma = (A, B, C, x^0)$ of dimension n define the mapping $\Phi_\Sigma : X_\Omega \to \mathbb{R}^p$ by $\Phi_\Sigma(G\omega) = y(1)$, where $y(1)$ is the solution of Σ with control $u = G\omega$. Similarly we define the mapping for sign-observations

$$S_\Sigma : X_\Omega \to \{0,1\}^p \quad G\omega \mapsto \operatorname{sign}(\Phi_\Sigma(G\omega)).$$

The class of above mappings is the *sign system concept class*, $\mathcal{F}_{m,p} = \{S_\Sigma \;;\; \Sigma \text{ linear system of dimension } n\}$.

By showing that the output of the control system is a sign of a rational expression and applying the Goldberg-Jerrum bound, we obtain:

t:vcub

Theorem 2.1.7. (VC-dimension upper bound, p = 1)

$$VC(\mathcal{F}_{m,1}) \leq 2(2mn^2 + 4n + 1)\log_2\Big[8e\big(m2n^2k4(n + \ell_{\max}) + 1\big)\big(6n^2k + 2k + 1\big)\Big],$$

where $\ell_{\max}$ is given by (2.2.3).

In terms of n (the dimension of the state space) and k (the band-width) the upper bound is of the form $O(n^2\log_2(nk))$.

Remark 2.1.13 indicates how the bound is formed when the input functions satisfy the more abstract rationality condition. In that case the degrees of the polynomials and the number of polynomial evaluations are different. However, in terms of n and k, the bound is of the same form. All VC-dimension or pseudo-dimension bounds stated in this paper can be modified for the rationality condition in the same way.

The lower bounds are in terms of n and k. They hold for linearly independent continuous control functions and compared to upper bounds, no particular form of the functions is needed. The bounds are obtained by imposing a specific structure on control systems, and a lower bound for a restricted class of control systems provides a lower bound for more general classes.

t:vclb

Theorem 2.1.8. (VC-dimension lower bounds, p = 1)

$$VC(\mathcal{F}_{1,1}) \geq \max\left\{ m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor, m' \right\}$$

where $m' = \min\{n, k\}$.

In terms of k the upper and the lower bound match up to a constant. For n and k the lower bound is of the form $O(n \log_2(k/n))$. Note that if the system dimension n is small compared to the band-width k the VC-dimension upper and lower bounds in Theorems 2.1.7 and 2.1.8 become tighter, both being of the form $c \log_2 k$ (with different values of the constant c).

Extending the upper bounds to the case of vector-valued observations can be done in various ways based on the result obtained for scalar observations. For example, we may consider the p-dimensional output as bits representing a number in $\{0, \ldots, 2^p - 1\}$ and introduce a loss function for each $f \in \mathcal{F}_{m,p}$ as $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. We define the VC-dimension of the p-dimensional observation as the VC-dimension of the above class of loss functions. Modifying the argument using the Goldberg-Jerrum bound leads to a bound of the following form:

t:vcub-p

Theorem 2.1.9. (VC-dimension upper bound)

$$\begin{aligned} VC(\mathcal{F}_{m,p}) &\leq 2(2pmn^2 + 4n + p) \\ &\quad \times \log_2\left[8e\left(m2n^2k4(n + \ell_{\max}) + 1\right)\left(2^p - 1 + p(2k+1) + 6n^2k\right)\right], \end{aligned}$$

where $\ell_{\max}$ *is given by* (2.2.3).

Next we state the main result concerning learnability of the actual input-output mapping, i.e., learning without taking the sign of the final state observation.

Definition 3. (Control system concept class, $\mathcal{G}_{p,L}$) Let $\widetilde{\mathcal{F}} = \{\Phi_\Sigma : X_\Omega \to \mathbb{R}^p \,;\, \Sigma \text{ linear system of dimension } n\}$ and define the *control system concept class* as $\mathcal{G}_{p,L} = A_{\widetilde{\mathcal{F}},L}$, where $A_{\widetilde{\mathcal{F}},L}$ is given by (2.2.4).

Methods for calculating upper bounds for the VC-dimension readily give a tool for obtaining upper bounds for the pseudo-dimension with respect to loss functions that preserve the rationality structure of the output. A typical example is illustrated by the loss function

e:pdloss

$$L(z_1, z_2) = (z_1 - z_2)^2/(1 + (z_1 - z_2)^2) \tag{2.2.7}$$

and the following result:

t:pdub

Theorem 2.1.10. (Pseudo-dimension upper bound, p = 1)

$$PD(\mathcal{G}_{1,L}) \leq 2\left(2mn^2 + 4n + 1\right)\log_2\Big[8e2\big(m2n^2k4(n + \ell_{\max}) + 1\big)\big(6n^2k + 2k + 1\big)\Big],$$

where the loss function L is given by (2.2.7) *and $\ell_{\max}$ is given by* (2.2.3).

This differs from the corresponding VC-dimension bound only by the maximum degree of the polynomials, which is doubled. Extending this pseudo-dimension bound for p-dimensional observations can be done naturally by modifying the loss function. Lower bounds for the VC-dimension are lower bounds for the pseudo-dimension as such.

We have seen how the rationality condition together with the Goldberg-Jerrum bound yields several upper bounds for the VC- or pseudo-dimension associated to linear continuous-time control systems.

2.1.2 An Upper Bound for the VC-Dimension with Scalar Observations

s:vcub-p1

We begin this section by proving Lemma 2.1.6. After that, as a lemma, we develop an upper bound for the VC-dimension induced by the control system (2.2.6) with its initial state fixed to be zero. Theorem 2.1.7 with an arbitrary initial condition is then a simple modification of the argument.

Proof of Lemma 2.1.6. If $\omega(t) = t^{\ell}e^{\alpha t}\sin(\beta t)$ or $\omega(t) = t^{\ell}e^{\alpha t}\cos(\beta t)$ with $\ell \leq \ell_{\max}$ then in place of

$$\int_0^1 t^{\tilde{\ell}}e^{at}\sin(bt)\omega(t)dt \quad \text{or} \quad \int_0^1 t^{\tilde{\ell}}e^{at}\cos(bt)\omega(t)dt$$

by combining exponents and using sum formulae for sin and cos (see Example 2.1.5) it is enough to study terms of the form

$$\int_0^1 t^{\tilde{k}}e^{\tilde{a}t}\sin(\tilde{b}t)dt \quad \text{or} \quad \int_0^1 t^{\tilde{k}}e^{\tilde{a}t}\cos(\tilde{b}t)dt,$$

where $\tilde{k} \in \{0, \ldots, n + \ell_{\max} - 1\}$. Because for $\tilde{k} > 0$

e:sinform

$$\begin{aligned}\int_0^1 t^{\tilde{k}}e^{\tilde{a}t}\sin(\tilde{b}t)dt = {} & \frac{e^{\tilde{a}}}{\tilde{a}^2+\tilde{b}^2}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) \\ & - \frac{\tilde{k}}{\tilde{a}^2+\tilde{b}^2}\int_0^1 t^{\tilde{k}-1}e^{\tilde{a}t}(\tilde{a}\sin(\tilde{b}t) - \tilde{b}\cos(\tilde{b}t))dt \end{aligned} \tag{2.2.8}$$

and

e:sinform2

$$\int_0^1 e^{\tilde{a}t}\sin(\tilde{b}t)dt = \frac{e^{\tilde{a}}}{\tilde{a}^2+\tilde{b}^2}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b} \tag{2.2.9}$$

and a similar formulae for $\int_0^1 t^{\tilde{k}}e^{\tilde{a}t}\cos(\tilde{b}t)dt$ holds, we see by induction that (RAT) is satisfied. When $\tilde{a} = \tilde{b} = 0$ the rationality conditon in clearly satisfied. By induction the numerator of $\int_0^1 t^{\tilde{k}}e^{\tilde{a}t}\sin(\tilde{b}t)dt$ is a polynomial of $\tilde{a}, \tilde{b}, e^{\tilde{a}}\cos\tilde{b}$ and $e^{\tilde{a}}\sin\tilde{b}$. By using sum formulae for sin and cos, the previous expression is in turn a polynomial of $a, b, e^a\cos b$ and $e^a\sin b$ because $\tilde{a} = a + \alpha$ and $\tilde{b} = b \pm \beta$ for some fixed α and β. By similar arguments, the denominator is a polynomial of a and b.

Note that, for example, e^{α} equals to a constant times e^a, so this process does not change the degrees of the polynomials. Further, observe that the denominator of

$\int_0^1 t^{\tilde{\ell}} e^{at} \sin(bt)\omega(t)dt$ consists of at most two products of variables a and b of the form $((a+\alpha)^2 + (b \pm \beta)^2)^{\tilde{\ell}+\ell+1}$. These terms are either zero at one point or positive. Thus there are at most two points where the denominator vanishes and h, the maximum number of pieces in the rationality condition, is at most three.

Finally, $d_{\max}$, the highest degree of polynomials appearing in (RAT), is at most $4(n + \ell_{\max})$. Equations (2.2.8) and (2.2.9) show that the degree of the numerator is not higher than the one of the denominator, which is at most $4(n + \ell_{\max})$. □

p:param

Proposition 2.1.11. *Denote the control functions by $\omega = (\omega_1, \ldots, \omega_k)^T$ and assume that each ω_i, $i = 1, \ldots, k$ satisfies the rationality condition (RAT) and let $\Lambda = \mathbb{R}^{2pn^2m} \times \mathbb{R}^{4n} \times \mathbb{R}^p$. Then there exists a mapping $H : \Lambda \times \mathbb{R}^{mk} \to \mathbb{R}^p$ (depending on ω) such that for each $\Sigma = (A, B, C, x^0)$ there exists a $\lambda \in \Lambda$ satistying*

$$\Phi_\Sigma(G\omega) = H(\lambda, G) \text{ for all } G \in \mathbb{R}^{mk}.$$

Proof. For system $\Sigma = (A, B, C, x^0)$,

$$\Phi_\Sigma(u) = y(1) = Ce^A x^0 + C \int_0^1 e^{A(1-t)} Bu(t)dt.$$

By an argument based on the real Jordan form of e^{At}, the entries of $e^{A(1-t)}$ are linear combinations of functions of the form $t^{\tilde{\ell}} e^{at} \cos(bt)$ and $t^{\tilde{\ell}} e^{at} \sin(bt)$, where $\tilde{\ell} \in \{0, \ldots, n-1\}$ and $a + ib$ is an eigenvalue of A. Hence we define the following $2n$ functions

$$\begin{aligned}
\xi_1(a, b, t) &= e^{at} \cos(bt) \\
\xi_2(a, b, t) &= t^2 e^{at} \cos(bt) \\
&\vdots \\
\xi_n(a, b, t) &= t^{n-1} e^{at} \cos(bt) \\
\xi_{n+1}(a, b, t) &= e^{at} \sin(bt) \\
&\vdots \\
\xi_{2n}(a, b, t) &= t^{n-1} e^{at} \sin(bt).
\end{aligned}$$

By the rationality condition (RAT)

$$\int_0^1 \xi_\ell(t)\omega_j(t)dt = \frac{\hat{P}_{\ell j}(a, b, e^a \cos b, e^a \sin b)}{\hat{Q}_{\ell j}(a, b, e^a \cos b, e^a \sin b)}$$

for all $a, b \in \mathbb{R}$ and where $\hat{P}_{\ell j}$ and $\hat{Q}_{\ell j}$ are piecewise polynomial expressions.

Let $H(\mathbf{A}, \mathbf{X}, h, G) = (H_1, \ldots, H_p)^T$, where for $1 \le \kappa \le p$

$$H_\kappa(\mathbf{A}, \mathbf{X}, h, G) = \sum_{i=1}^{m} \sum_{\beta=1}^{n} \sum_{\ell=1}^{2n} \alpha_{i\beta\ell\kappa} \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}(a, b, e^a \cos b, e^a \sin b)}{\hat{Q}_{\ell j}(a, b, e^a \cos b, e^a \sin b)} + h_\kappa$$

and

$$\mathbf{A} = (\alpha_{i\beta\ell\kappa})_{\substack{i=1,\ldots,m \\ \beta=1,\ldots,n \\ \ell=1,\ldots,2n \\ \kappa=1,\ldots,p}} \qquad \mathbf{X} = (x_{\beta\eta})_{\substack{\beta=1,\ldots,n \\ \eta=1,\ldots,4}}$$

$$h = (h_1, \ldots, h_p)^T \qquad G = (g_{ij})_{\substack{i=1,\ldots,m \\ j=1,\ldots,k}}.$$

Next we relate Φ_Σ and H and we write

$$Ce^{A(1-t)}B = \begin{bmatrix} \gamma_{11} & \cdots & \gamma_{1m} \\ \vdots & & \vdots \\ \gamma_{p1} & \cdots & \gamma_{pm} \end{bmatrix}.$$

We list the eigenvalues of A as $a_\beta + ib_\beta$ for $\beta = 1, \ldots, n$ and let $\xi_{\beta\ell}(t) = \xi_\ell(a_\beta, b_\beta, t)$ for $\beta = 1, \ldots, n$ and $\ell = 1, \ldots, 2n$. Then there exists some $(\alpha_{i\beta\ell\kappa})$ such that

e:alp-par

$$\gamma_{\kappa i}(t) = \sum_{\beta=1}^{n} \sum_{\ell=1}^{2n} \alpha_{i\beta\ell\kappa}\xi_{\beta\kappa}(t). \tag{2.2.10}$$

Let $\lambda = (\mathbf{A}, \mathbf{X}, h)$, where $\mathbf{A}$ satisfies (2.2.10), $\mathbf{X} = (x_{\beta\eta})$, where $x_{\beta 1} = a_\beta, x_{\beta 2} = b_\beta, x_{\beta 3} = e^{a_\beta} \cos b_\beta$ and $x_{\beta 4} = e^{a_\beta} \sin b_\beta$, and $h = Ce^A x^0$. We claim that

$$H(\lambda, G) = y(1) = \Phi_\Sigma(G\omega) \text{ for all } G \in \mathbb{R}^{mk}.$$

Note that the κ-th component of $\Phi_\Sigma(G\omega)$ is given by

$$
\begin{aligned}
&\int_0^1 \sum_{i=1}^{m} \gamma_{\kappa i}(t) u_i(t)\,dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{\beta=1}^{n}\sum_{\ell=1}^{2n} \alpha_{i\beta\ell\kappa} \int_0^1 \xi_\ell(t) \sum_{j=1}^{k} g_{ij}\omega_j(t)\,dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{\beta=1}^{n}\sum_{\ell=1}^{2n} \alpha_{i\beta\ell\kappa} \sum_{j=1}^{k} g_{ij} \int_0^1 \xi_\ell(t)\omega_j(t)\,dt + h_\kappa \\
&= \sum_{i=1}^{m}\sum_{\beta=1}^{n}\sum_{\ell=1}^{2n} \alpha_{i\beta\ell\kappa} \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}(x_{\beta 1}, x_{\beta 2}, x_{\beta 3}, x_{\beta 4})}{\hat{Q}_{\ell j}((x_{\beta 1}, x_{\beta 2}, x_{\beta 3}, x_{\beta 4})} + h_\kappa \\
&= H_\kappa(\mathbf{A}, \mathbf{X}, h, G).
\end{aligned}
$$

□

Next we take $p = 1$ and study the VC-dimension of the sign control concept class, $\mathcal{F}_{m,1}$, where each control parametrized by G gives rise to $\operatorname{sign} y(1)$.

l:vcub

Lemma 2.1.12. *$\mathcal{F}_{m,1}$ with initial condition $x(0) = 0$ satisfies*

$$VC(\mathcal{F}_{m,1}) \leq 2\left(2mn^2 + 4n\right) \log_2\Big[8e\big(m2n^2k4(n + \ell_{\max}) + 1\big)\big(6n^2k + 2k + 1\big)\Big],$$

where $\ell_{\max}$ is given by (2.2.3).

Proof. By Proposition 2.1.11 $y(1) = H(\lambda, G)$ where $\lambda \in \mathbb{R}^{2mn^2} \times \mathbb{R}^{4n}$ are considered as parameters. In fact, $y(1) = \frac{\mathbf{P}}{\mathbf{Q}}$, where $\mathbf{P}$ and $\mathbf{Q}$ denote piecewise polynomial functions. As in the statement of Goldberg-Jerrum bounds we have a function $F : \Lambda \times \mathbb{R}^{mk} \to \{0, 1\}$ defined by $F(\lambda, G) = \operatorname{sign} H(\lambda, G)$. The concept class associated to the system identification problem is $\mathcal{F} := \{F(\lambda, \cdot) : \mathbb{R}^{mk} \to \{0, 1\}\ ;\ \lambda \in \Lambda\}$, where $\Lambda = \mathbb{R}^{2mn^2+4n}$. Before applying the Goldberg-Jerrum bound we need to determine the possible degrees of $\mathbf{P}$ and $\mathbf{Q}$ with respect to the parameters.

The rationality condition implies that

$$\max_{\substack{i \leq \ell \leq n \\ 1 \leq j \leq k}} \left\{\deg(\hat{P}_{\ell j}), \deg(\hat{Q}_{\ell j})\right\} \leq d_{\max}.$$

Then

$$\frac{\widetilde{P}_{i\ell}}{\widetilde{Q}_{i\ell}} = \sum_{j=1}^{k} g_{ij} \frac{\hat{P}_{\ell j}}{\hat{Q}_{\ell j}}$$

so $\deg(\widetilde{Q}_{i\ell}) \leq kd_{\max}$ and $\deg(\widetilde{P}_{i\ell}) \leq kd_{\max}$. Note here that we are calculating the degree with respect to the system parameters, and the inputs g_{ij} do not contribute. By continuing in a similar fashion and combining β-summation to the ℓ-summation in Proposition 2.1.11, we write $P_i/Q_i = \sum_{\ell=1}^{2n^2} \alpha_{i\ell}\widetilde{P}_{i\ell}/\widetilde{Q}_{i\ell}$ to conclude that $\deg(Q_i) \leq 2n^2 \deg(\widetilde{Q}_{i\ell}) = 2n^2 kd_{\max}$ and $\deg(P_i) \leq 2n^2 kd_{\max} + 1$. Finally, $\mathbf{P}/\mathbf{Q} = \sum_{i=1}^{m} P_i/Q_i$ with $\deg(\mathbf{Q}) \leq m2n^2kd_{\max}$ and $\deg(\mathbf{P}) \leq m2n^2kd_{\max} + 1$.

Recall that for each i and j, $i \in \{1, \ldots, 2n^2\}$, $j \in \{1, \ldots, k\}$ the denominator of $\int \xi_i(t)\omega_j(t)dt$ can vanish at most at 2 points (when A has complex eigenvalues) determined by the control function ω_j. Hence the number of different $\frac{\mathbf{P}}{\mathbf{Q}}$ expressions is at most $2k+1$. Which expression the Boolean formula will "look at" is determined by $2n^2kh$ polynomial evaluations. Using Lemma 2.1.6 for bounds on h and $d_{\max}$, we apply the Goldberg-Jerrum bounds with $s = 6n^2k+2k+1$, $d = m2n^2k4(n+\ell_{\max})+1$ and $\ell = 2mn^2 + 4n$. $\square$

A simple example of the piecewise polynomial function $\mathbf{P}/\mathbf{Q}$ together with the decision table for the final output is provided in Appendix A.

r:ratbound

Remark 2.1.13. The VC-dimension bound is modified for the more abstract rationality conditions as follows. Evaluating the sign of the output involves the evaluation of $2(8ed_{\max}2n^2kh/4)^4 + 2n^2kh$ polynomials; $2n^2kh$ evaluations are needed to find an appropriate piece and by Theorem 2.1.3 the maximum number of possible expressions of the type $\mathbf{P}/\mathbf{Q}$ is bounded by $(8ed_{\max}2n^2kh/4)^4$. Applying the Goldberg-Jerrum bound with $s = 2(8ed_{\max}2n^2kh/4)^4 + 2n^2kh$, $d = m2n^2kd_{\max} + 1$ and $\ell = 2mn^2 + 4n$ gives the result.

Proof of the VC-dimension upper bound, $\mathbf{p = 1}$. By using the previous notation $y = Ce^A x^0 + C\int_0^1 e^{A(1-t)}Bu(t)dt$. Let $\widetilde{x} = Ce^A x^0$. Then $y = \widetilde{x} + \mathbf{P}/\mathbf{Q} =$

$(\widetilde{x}\mathbf{Q}+\mathbf{P})/\mathbf{Q} = \widetilde{\mathbf{P}}/\mathbf{Q}$. This has $2mn^2 + 4n + 1$ parameters and $\deg(\widetilde{\mathbf{P}}) \leq m2n^2kd_{\max} + 1$. □

Remark 2.1.14. Taking $2n^2$ functions $\xi_{\beta\ell}(t)$, $\beta = 1, \ldots, n$, $\ell = 1, \ldots, 2n$ in Proposition 2.1.11 is clear overcounting. Preliminary calculations suggest that $O(n \ln n)$ functions $\xi_{\beta\ell}(t)$ would be enough.

2.1.3 Lower Bounds for the VC-Dimension

The lower bounds for the VC-dimension are developed for a single-input single-output system with initial state zero. The control is

$$u = \sum_{j=1}^{k} g_j \omega_j : [0,1] \to \mathbb{R}.$$

We derive lower bounds by fixing the structure of A, B, and C, and using the dual VC-dimension and axis shattering following the ideas of DasGupta and Sontag [14]. Lemmata in this section prove Theorem 2.1.8. These lower bounds are very general; we just assume that the input functions are continuous and linearly independent, thus no particular structure is required as in the upper bounds.

To make the next proof cleaner we formulate a part of it as a separate proposition.

p:linind

Proposition 2.1.15. *Let $\omega_j : [0,1] \to \mathbb{R}$, $j = 1, \ldots, k$ be continuous and linearly independent. Then the functions*

$$h_j(\lambda) = \int_0^1 e^{\lambda t} \omega_j(t) dt, \quad j = 1, \ldots, k$$

are linearly independent.

Proof. We need to show that for any $f \in C[0,1] = \{f\colon [0,1] \to \mathbb{R}\ ;\ f \text{ continuous}\}$

$$\int_0^1 e^{\lambda t} f(t) dt = 0 \qquad \text{for all } \lambda \in \mathbb{R}$$

implies $f(t) \equiv 0$ on $[0,1]$. We can write

$$\int_0^1 e^{\lambda t} f(t)dt = \int_0^1 \sum_{k=1}^{\infty} \frac{1}{k!}(\lambda t)^k f(t)dt.$$

Because

$$\sum_{k=1}^{\infty} \frac{1}{k!}\lambda^k \int_0^1 t^k |f(t)|dt \leq \sum_{k=1}^{\infty} \frac{1}{k!}\lambda^k \sup_{t\in[0,1]} |f(t)| \leq \sup_{t\in[0,1]} |f(t)|e^{\lambda} < \infty,$$

the order of integration and summation can be changed by Fubini's Theorem. Thus

$$\int_0^1 e^{\lambda t} f(t)dt = \sum_{k=1}^{\infty} \frac{1}{k!}\lambda^k \int_0^1 t^k f(t)dt = 0$$

for all $\lambda \in \mathbb{R}$. This further implies that $\int_0^1 t^k f(t)dt = 0$ for all $k = 1, 2, \ldots$. As f is continuous and polynomials are dense in $C[0,1]$, we have that $f(t) \equiv 0$.

Linear independence of $h_1, \ldots, h_k$ follows: If $\alpha_1 h_1(\lambda) + \cdots + \alpha_k h_k(\lambda) = 0$ for all $\lambda \in \mathbb{R}$, then

$$\int_0^1 e^{\lambda t}(\alpha_1\omega_1(t) + \cdots + \alpha_k\omega_k(t))dt = 0, \quad \forall \lambda \in \mathbb{R}$$

and by the above argument $\alpha_1\omega_1(t) + \cdots + \alpha_k\omega_k(t) = 0$ for all t. As $\omega_1, \ldots, \omega_k$ were assumed to be linearly independent, we have that $(\alpha_1, \ldots, \alpha_k) = (0, \ldots, 0)$. □

l:vclb1

Lemma 2.1.16 (Lower bound 1).

$$VC(\mathcal{F}_{1,1}) \geq m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor,$$

where $m' = \min\{n, k\}$.

Proof. Let $\omega_j(t)$, $j = 1, \ldots, k$ be continuous and linearly independent. Let A have n distinct real eigenvalues $-\lambda_1, \ldots, -\lambda_n$, and take B and C so that

$$Ce^{A(1-t)}B = \sum_{i=1}^{m'} e^{\lambda_i t},$$

where $m' = \min\{n, k\}$. Then the final output of the system is

$$y(1) = \int_0^1 Ce^{A(1-t)}B \sum_{j=1}^{k} g_j\omega_j(t)dt = \sum_{i=1}^{m'} \sum_{j=1}^{k} g_j \int_0^1 e^{\lambda_i t}\omega_j(t)dt.$$

Define $h_j(\lambda) = \int_0^1 e^{\lambda t}\omega_j(t)dt$. By Proposition 2.1.15 the h_j's are linearly independent and we can find $\lambda_1, \ldots, \lambda_k$ such that the matrix

$$\begin{bmatrix} h_i(\lambda_i) & \cdots & h_k(\lambda_k) \\ \vdots & & \vdots \\ h_1(\lambda_k) & \cdots & h_k(\lambda_k) \end{bmatrix}$$

has rank k.

The control system with sign observations gives the mapping $F : \mathbb{R}^{m'} \times \mathbb{R}^k \to \{0,1\}$ by

$$(\lambda_1, \ldots, \lambda_{m'}, g_1, \ldots, g_k) \mapsto \operatorname{sign}\left[\sum_{i=1}^{m'}\sum_{j=1}^{k} g_j h_j(\lambda_i)\right].$$

We show that the mapping from parameters $\lambda_1, \ldots, \lambda_{m'}$ to $\{0,1\}$ can be axis shattered. Let $L = \{\lambda_1, \ldots, \lambda_k\}$ be so that $[h_j(\lambda_i)]_{i,j}$ has rank k. Denote by $L_1, \ldots, L_{m'}$ disjoint subsets of L such that $|L_i| = \lfloor k/m' \rfloor$ and let $M = L \setminus \{\bigcup_{i=1}^{m'} L_i\}$. Next we want to interpolate in the points of L.

Fix s, $1 \le s \le m'$ and let $\phi : L_s \to \{0,1\}$ be any dichotomy. Next find $g_1, \ldots, g_k$ such that

e:dicho

$$\begin{aligned} \sum_{j=1}^{k} g_j h_j(\lambda_s) &= \phi(l_s), \qquad \forall\, \lambda_s \in L_s, \\ \sum_{j=1}^{k} g_j h_j(\lambda) &= 0, \qquad \forall\, \lambda \in (L \cup M) \setminus L_s. \end{aligned} \tag{2.2.11}$$

Let $g_1^*, \ldots g_k^*$ satisfy (2.2.11). (A unique solution exists because $[h_j(\lambda_i)]$ has rank k.) Then

$$F[\lambda_1, \ldots, \lambda_{m'}, g_1^*, \ldots, g_k^*) = \operatorname{sign}\left[\sum_{i=1}^{m'}\sum_{j=1}^{k} g_j^* h_j(\lambda_i)\right] = \phi(\lambda),$$

when $\lambda \in L_s$ and for all $(\lambda_1, \ldots, \lambda_{m'}) \in L_1 \times \cdots \times L_{m'}$.

Let $\widetilde{\mathcal{F}} = \{F(\lambda_1, \ldots, \lambda_{m'}, \cdot) : \mathbb{R}^k \to \{0,1\}\ ;\ (\lambda_1, \ldots, \lambda_{m'}) \in \mathbb{R}^{m'}\}$. By the Axis shattering bound given in Theorem 2.1.1

$$\mathrm{VC}(\widetilde{\mathcal{F}}) \ge m' \left\lfloor \log_2 \left\lfloor \frac{k}{m'} \right\rfloor \right\rfloor$$

and thus $\mathrm{VC}(\mathcal{F}_{1,1}) \geq \mathrm{VC}(\widetilde{\mathcal{F}})$, where $\mathcal{F}_{1,1}$ is the control system concept class with $p = m = 1$. $\square$

l:vclb2

Lemma 2.1.17 (Lower bound 2). *If $k \leq n$ then*

$$VC(\mathcal{F}_{1,1}) \geq k.$$

Proof. We make a small modification of the above argument. Assume that $k \leq n$ and let A have n real eigenvalues $\lambda_1, \ldots, \lambda_n$. Next we take B and C so that $Ce^{A(1-t)}B = \sum_{i=1}^n e^{\lambda_i t}\beta_i$, where $(\beta_1, \ldots, \beta_n, \lambda_1, \ldots, \lambda_n)$ are considered as system parameters.

We study the mapping

$$\begin{aligned}(\beta_1, \ldots, \beta_n, \lambda_1, \ldots, \lambda_n, g_1, \ldots, g_k) \mapsto \operatorname{sign}&\left[\sum_{i=1}^n \sum_{j=1}^k g_j h_j(\lambda_i)\beta_i\right] \\ &= \operatorname{sign}\left[\sum_{j=1}^k g_j \underbrace{\sum_{i=1}^n h_j(\lambda_i)\beta_i}_{\gamma_j}\right] = \operatorname{sign}\left[\sum_{j=1}^k g_j\gamma_j\right].\end{aligned}$$

Given $(\gamma_1, \ldots, \gamma_k)$, by linear independence of $h_1, \ldots, h_k$, we can find $\lambda_1, \ldots, \lambda_n, \beta_1, \ldots, \beta_n$ such that $\sum_{i=1}^n h_j(\lambda_i)\beta_i = \gamma_j$, $j = 1, \ldots, k$. But $(\gamma_1, \ldots, \gamma_k)$ can be viewed as a normal vector for a hyperplane through origin in $\mathbb{R}^k$ and the concept class associated to the mapping $(g_1, \ldots, g_k) \mapsto \operatorname{sign}\left[\sum_{j=1}^k g_j\gamma_j\right]$ as $(\gamma_1, \ldots, \gamma_k)$ varies has VC-dimension k. Hence $\mathrm{VC}(\mathcal{F}_{1,1}) \geq k$. $\square$

l:vclb3

Lemma 2.1.18 (Lower bound 3). *If $n \leq k$ then*

$$VC(\mathcal{F}_{1,1}) \geq n.$$

Proof. Our construction for the control system is as in the previous proof but now we assume that $n \leq k$ and we study

$$\begin{aligned}(\beta_1, \ldots, \beta_n, \lambda_1, \ldots, \lambda_n, g_1, \ldots, g_k) \mapsto \operatorname{sign}\left[\sum_{i=1}^{n}\sum_{j=1}^{k} g_j h_j(\lambda_i)\beta_i\right] \\ = \operatorname{sign}\left[\sum_{i=1}^{n}\underbrace{\sum_{j=1}^{k} g_j h_j(\lambda_i)}_{\tilde{g}_i}\beta_i\right] = \operatorname{sign}\left[\sum_{i=1}^{n}\tilde{g}_i\beta_i\right]\end{aligned}$$

and again by linear independence and the above hyperplane argument (now via first transforming $(g_1, \ldots, g_k)$) we can conclude that the above mapping has VC-dimension n. Thus $\mathrm{VC}(\mathcal{F}_{1,1}) \geq n$. □

Lemmata 2.1.16, 2.1.17 and 2.1.18 together prove Theorem 2.1.8.

2.1.4 VC-Dimension Upper Bounds for p-Dimensional Outputs

We begin by proving Theorem 2.1.9.

Proof of the VC-dimension upper bound. We develop an upper bound based on the bound for a scalar sign-observation. We have seen that under the rationality assumption (RAT) the scalar output is a piecewise rational expression $\mathbf{P}/\mathbf{Q}$. In general, the control system maps G to $(\operatorname{sign}(\mathbf{P}_1/\mathbf{Q}_1), \ldots, \operatorname{sign}(\mathbf{P}_p/\mathbf{Q}_p))^T$, which is understood as a binary representation of a number in $\{0, 1, \ldots, 2^p - 1\}$. Let $f : \mathbb{R}^{mk} \to \{0, \ldots, 2^p - 1\}$ be the mapping given by the control system, and denote the class of all such mappings by $\mathcal{F}$. For each $f \in \mathcal{F}$ introduce a loss function $L_{0\text{-}1,f}(z, a) = L_{0\text{-}1}(f(z), a) = 1$, when $f(z) \neq a$, and 0 otherwise. Define the class $L_{0\text{-}1,\mathcal{F}} = \{L_{0\text{-}1,f}\ ;\ f \in \mathcal{F}\}$.

In order to calculate the value of the output, after determining an appropriate piece, one needs to know the truth values of the expressions $\mathbf{P}_1 > 0, \mathbf{Q}_1 > 0, \ldots, \mathbf{P}_p > 0$ and $\mathbf{Q}_p > 0$, where $\mathbf{P}$'s and $\mathbf{Q}$'s are polynomials on inputs and parameters of the control system. To evaluate the value of the loss function $L_{0\text{-}1,f}(z,a)$, one needs the truth values of $y = 0, y = 1, \ldots, y = 2^p - 2$.

In the general case one needs $6n^2k + p(2k+1) + 2^p - 1$ truth values. As this procedure evaluates only polynomials, we can use the Goldberg-Jerrum bound again. The maximum degree of the polynomials is $m2n^2k4(n + \ell_{\max}) + 1$, and the total number of parameters is $2pn^2m + 4n + p$, where the last term comes from the initial condition. □

Remark 2.1.19. An upper bound can be derived by using the uniform Ψ-dimension defined by Ben-David et. al. [10]. As before let $f : \mathbb{R}^{mk} \to \{0, \ldots, 2^p - 1\}$ be the mapping given by the control system. Let $\Psi = \{\psi^1, \ldots, \psi^p\}$ be a collection of distinguishers $\psi^j : \{0, \ldots, 2^p - 1\} \to \{0, 1\}$ given by $\psi^j(\mathrm{sign}(y_1), \ldots, \mathrm{sign}(y_p)) = \mathrm{sign}(y_j)$.

Let $\psi^j_f : \mathbb{R}^{mk} \to \{0, 1\}$ be given by $\psi^j_f(z) = \psi^j(f(z))$ and $\psi^j_{\mathcal{F}} = \{\psi^j_f \,;\, f \in \mathcal{F}\}$. An upper bound for $\mathrm{VC}(\psi^j_{\mathcal{F}})$ is given by Theorem 2.1.7. Ben-David et. al. define the uniform Ψ-dimension of $\mathcal{F}$ in which only one distinguisher is selected to shatter

$$\Psi - \dim_U(\mathcal{F}) = \max\{\mathrm{VC}(\psi^j_{\mathcal{F}}) \,;\, \psi^j \in \Psi\}.$$

That is, the upper bound for the uniform Ψ-dimension is just the upper bound obtained earlier for scalar observations. However, the uniform Ψ-dimension can not be compared to the VC-dimension as such. For example, the formula for sample complexity with Ψ-dimension is different: to calculate the sample complexity the uniform Ψ-dimension is multiplied by $p \log_2(2^p - 1)$, see [10, Theorem 27].

2.1.5 Pseudo-Dimension Upper Bounds

In this section we illustrate a technique for modifying the VC-dimension bounds in order to derive bounds for the pseudo-dimension. This technique can be used with loss functions that preserve the rationality structure.

Next we prove Theorem 2.1.10.

Proof of the pseudo-dimension upper bound, $\mathbf{p = 1}$. The proof of Lemma 2.1.12 shows that the output at time 1 can be written as $y(1) = \mathbf{P}/\mathbf{Q}$, where $\mathbf{P}$ and $\mathbf{Q}$ are polynomials defined in pieces. Hence by using the loss function (2.2.7),

$$L(y(1), y) - \rho = \frac{(\mathbf{P}/\mathbf{Q} - y)^2}{1 + (\mathbf{P}/\mathbf{Q} - y)^2} - \rho = \frac{(\mathbf{P} - y\mathbf{Q})^2 - \rho(\mathbf{Q} - (\mathbf{P} - y\mathbf{Q})^2)}{\mathbf{Q}^2 - (\mathbf{P} - y\mathbf{Q})^2} = \frac{\hat{\mathbf{P}}}{\hat{\mathbf{Q}}}.$$

Applying the Goldberg-Jerrum bound by changing the maximum degree of the polynomials to be $2(m2n^2kd_{\max} + 1)$ gives an upper bound for the pseudo-dimension. Note that the Goldberg-Jerrum bound (notation refers to the statement of Theorem 2.1.2) does not depend on the dimension of the space X. It only depends on the dimension of the parameter space Λ, which remains the same for pseudo-dimension bounds. □

Remark 2.1.20. Different types of loss functions that preserve the rationality structure can be used with minor modifications. If the loss function is defined by using absolute values it increases the number of polynomial evaluations involved.

Remark 2.1.21. For p-dimensional outputs define a loss function $L : \mathbb{R}^p \times \mathbb{R}^p \to [0, 1]$ as

$$L(z, y) = \frac{\sum_{i=1}^{p}(z_i - y_i)^2}{1 + \sum_{i=1}^{p}(z_i - y_i)^2}.$$

The Goldberg-Jerrum argument with $s = 6n^2k + p(2k + 1)$, $d = 2(m2n^2k4(n + \ell_{\max}) + 1)$, and $\ell = 2pmn^2 + 4n + p$ gives a pseudo-dimension upper bound. Note that compared to the upper bound with p-dimensional sign-observations, here we actually need fewer polynomial evaluations as we no longer evaluate the truth values of $y = 0, y = 1$ etc.

2.1.6 A Class of Systems with VC-Dimension k

s:boundk

For the control system (2.2.1) with scalar control $u(t) = \sum_{i=1}^{k} g_i\omega_i(t)$ and unrestricted $\omega_1, \ldots, \omega_k$, the standard half-space argument gives an upper bound k. This bound is tight. We will give an example of a single-input, single-output one parameter family of control systems in dimension two that has VC-dimension k, when the controls are of the form $u(t) = \sum_{i=1}^{k} g_i \omega_i(t)$ and $\omega_i(1-t) = 1_{[2^{-i}, 2^{-i}+2^{\alpha}]}$, where $\alpha = -2(k+1)$.

Consider a control system

e:syst-vck

$$
\begin{aligned}
\dot{x_1} &= x_2 \\
\dot{x_2} &= -\lambda^2 x_1 + u \\
y &= -x_1.
\end{aligned}
\tag{2.2.12}
$$

For time interval [0,1] and initial condition $(x_1, x_2) = (0,0)$ the output is given by $y(1) = \int_0^1 \sin(\lambda t) u(1-t) dt$.

l:vck

Lemma 2.1.22. *Controls* $\{\omega_1, \ldots, \omega_k\}$ *such that* $\omega_i(1-t) = 1_{[2^{-i}, 2^{-i}+2^{\alpha}]}$, *where* $\alpha = -2(k+1)$, *are shattered by the control system* (2.2.12) *with sign-observations.*

Proof. Let $T = \{2^{-i},\ i = 1, \ldots, k\}$ and $J \subseteq T$. Define $\lambda_J = \pi \sum_{i=1}^{k} a_i 2^i$, where $a_i = 1$ if $2^{-i} \notin J$ and $a_i = 0$ otherwise. Now if $t = 2^{-\ell}$, then

$$
\lambda_J t = \pi \sum_{i=1}^{k} a_i 2^{i-\ell} = \pi \left(\underbrace{\sum_{i=1}^{\ell-1} a_i 2^{i-\ell}}_{1/2\, c_2} + a_\ell + \underbrace{\sum_{i=\ell+1}^{k} a_i 2^{i-\ell}}_{2c_1} \right),
$$

where $c_1 \in \mathbb{N}$ and $0 \leq c_2 < 2$.

Hence $\sin(\lambda_J t) = \sin(\pi(1/2\, c_2 + a_\ell))$. Note that if $a_\ell = 0$ then $1/2\, c_2 + a_\ell \in [0, 1 - 2^{-\ell}]$ and if $a_\ell = 1$ then $1/2\, c_2 + a_\ell \in [1, 2 - 2^{-\ell}]$. Thus $\sin(\pi(1/2\, c_2 + a_\ell)) \geq 0$ if $a_\ell = 0$ and $\sin(\pi(1/2\, c_2 + a_\ell)) \leq 0$ if $a_\ell = 1$. Therefore,

$$
\sin(\lambda_J t) \geq 0 \ \Leftrightarrow\ a_\ell = 0
$$

and

$$
\sin(\lambda_J t) \geq 0 \ \Leftrightarrow\ t \in J.
$$

Further,

$$\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt \geq 0 \Leftrightarrow a_\ell = 0,$$

where α is taken so that

e:alpha

$$\sum_{j=1}^{k} 2^j 2^\alpha \leq 2^{-(k+1)}. \tag{2.2.13}$$

This assures that when $\ell \leq k$ and $t \in [2^{-\ell}, 2^{-\ell} + 2^\alpha]$

$$\lambda_J t \in [0, \pi(1 - 2^{-\ell} + \sum_{j=1}^{k} 2^j 2^\alpha)] \subset [0, \pi)\,, \text{ if } a_\ell = 0$$

or similarly

$$\lambda_J t \in [\pi, 2\pi)\,, \text{ if } a_\ell = 1.$$

In (2.2.13) we can take $\alpha = -2(k+1)$ as $\sum_{j=1}^{k} 2^j = 2^{k+1} - 2$.

In this way the integrand in

$$\int_{2^{-\ell}}^{2^{-\ell}+2^{\alpha}} \sin(\lambda_J t)dt$$

is either positive or negative.

For $S \subseteq \{1, \ldots, k\}$, let $J = \{2^{-i}, i \in S\}$. For each ω_i,

$$\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt = \int_{2^{-i}}^{2^{-i}+2^{\alpha}} \sin(\lambda_J t)dt > 0 \Leftrightarrow i \in S,$$

i.e., the set of controls $\{\omega_1, \ldots, \omega_k\}$ is shattered by the mapping

$$\omega_i \mapsto \text{sign}\left[\int_0^1 \sin(\lambda_J t)\omega_i(1-t)dt\right]$$

$\square$

2.1.7 An Application to Learning

We illustrate how previously obtained learning complexity bounds can be applied in the system identification. Let us study for simplicity scalar output systems

$$\dot{x} = Ax + Bu, \quad x(0) = x^0, \quad y = Cx,$$

$$u = G\omega, \quad \text{and} \quad \omega_j \in \Omega \text{ for } j = 1, \ldots, k.$$

with sign observations. Entries of A, B, x^0, and C are considered as control system parameters to be learned (though less parameters are actually needed) and we denote the space of control system parameters by Λ. Controls are, in turn, parametrized by the entries of G and the set Ω given by (2.2.2) is fixed.

Thus our concept class is the sign systems concept class $\mathcal{F}_{m,1}$ given in Definition 2. Recall that our VC-dimension bounds (and hence all other bounds derived from those) are in terms of system dimensions and constants related to controls. In particular, the VC-dimension depends on the system dimension n and in the system identification setting we should not assume that the exact system dimension is known. Instead, we assume that there is prior knowledge on the possible values of the system dimension and hence we consider a hierarchy of classes

$$\mathcal{H}_{n_1} \subset \mathcal{H}_{n_2} \subset \cdots \subset \mathcal{H}_{n_d}$$

so that $\mathcal{H}_{n_i}$ is a sign system concept class $\mathcal{F}_{m,1}$ with system dimension n_i. That is, the system dimension increases in the hierarcy while the other structure remains constant.

In the learning setting controls, parametrized by the matrix G, are generated at random according to some probability measure P, which may be unknown. Assume that the unknown target system T lies in some subclass $\mathcal{H}_{n_i}$. A fixed number of labeled samples $z = \{(G_1, T(G_1)), \ldots, (G_m, T(G_m))\}$, where $T(G_i)$ is the sign observation of the system with control given by G_i, is presented to the learner. Then the learner uses an algorithm to find a consistent hypothesis $h \in \mathcal{H}_{n_j}$ for some $1 \leq j \leq d$. Recall that a hypothesis is consistent if it agrees with the observed multisample z. The generalization error $\mathrm{er}_P(h) = P\{G \ ; \ h(G) \neq T(G)\}$ is the probability that h will misclassify a future unseen sample.

In addition to the standard setting above we need positive numbers p_{n_i} such that $\sum_{i=1}^{d} p_{n_i} = 1$. Each p_{n_i} represents our prior estimate for the probability that $\mathcal{H}_{n_i}$ is the smallest class containing a consistent hypothesis. Given a confidence parameter

δ, the following theorem provides a function $\epsilon(m, D_{n_i}, p_{n_i}, \delta)$ that gives an upper bound on the generalization error of h with confidence $1-\delta$.

Theorem 2.1.23 t:str-min ([34]). *Let $\mathcal{H}_{n_i}$, $i = 1, \ldots, d$ be a sequence of hypothesis classes mapping X to $\{0, 1\}$ such that $VC(\mathcal{H}_{n_i}) = D_{n_i}$ and let P be a probability distribution on X. Let p_{n_i}, $i = 1, \ldots, d$ be positive numbers satisfying $\sum_{i=1}^{d} p_{n_i} = 1$. With probability $1-\delta$ over m independent examples drawn according to P, for any n_j for which a learner finds a consistent hypothesis h in $\mathcal{H}_{n_j}$, the generalization error of h is bounded from above by*

$$\epsilon(m, D_{n_j}, p_{n_j}, \delta) = \frac{4}{m}\left(D_{n_j} \ln\left(\frac{2em}{D_{n_j}}\right) + \ln\left(\frac{1}{p_{n_j}}\right) + \ln\left(\frac{4}{\delta}\right)\right)$$

provided $D_{n_j} \leq m$.

In this setting Theorem 2.1.7 provides an upper bound for D_{n_j}, that is,

$$D_{n_j} = VC(\mathcal{H}_{n_j}) \leq 2(2mn_j^2 + 4n_j + 1)\log_2\left[8e\big(2mn_j^2k4(n_j + \ell_{\max}) + 1\big)(6n_j^2k + 2k + 1)\right].$$

In terms of sample complexities Theorem 2.1.23 gives the following corollary:

Corollary 2.1.24. *The generalization error of h is at most ϵ provided that the sample size is at least*

$$\frac{8D_{n_j}}{\epsilon}\ln\left(\frac{16}{\epsilon}\right) + 2\ln\left(\frac{4}{p_{n_j}\delta}\right).$$

Proof. $\epsilon(m, D_{n_j}, p_{n_j}, \delta) \leq \epsilon$ implies that

$$m \geq \frac{4}{\epsilon}\left(D_{n_j}\ln m + D_{n_j}\ln\left(\frac{2e}{D_{n_j}}\right) + \ln\left(\frac{4}{p_{n_j}\delta}\right)\right).$$

By using the fact that $\ln x \leq \alpha x - \ln\alpha - 1$ for all $\alpha, x > 0$ [5], we can bound

$$\begin{aligned}\frac{4D_{n_j}}{\epsilon}\ln m &\leq \frac{4D_{n_j}}{\epsilon}\left(\frac{\epsilon m}{8D_{n_j}} - \ln\left(\frac{\epsilon}{8D_{n_j}}\right) - 1\right)\\ &= \frac{m}{2} + \frac{4D_{n_j}}{\epsilon}\ln\left(\frac{8D_{n_j}}{e\epsilon}\right),\end{aligned}$$

and hence

$$m \geq \frac{m}{2} + \frac{4D_{n_j}}{\epsilon} \ln\left(\frac{8D_{n_j}}{e\epsilon}\right) + D_{n_j} \ln\left(\frac{2e}{D_{n_j}}\right) + \ln\left(\frac{4}{p_{n_j}\delta}\right)$$

and further

$$\frac{m}{2} \geq \frac{4D_{n_j}}{\epsilon} \ln\left(\frac{8D_{n_j}}{e\epsilon}\right) + \frac{4D_{n_j}}{\epsilon}\frac{\epsilon}{4} \ln\left(\frac{2e}{D_{n_j}}\right) + \ln\left(\frac{4}{p_{n_j}\delta}\right).$$

The above is satisfied if

$$m \geq \frac{8D_{n_j}}{\epsilon} \ln\left(\frac{16}{\epsilon}\right) + 2\ln\left(\frac{4}{p_{n_j}\delta}\right).$$

□

2.2 Fat-Shattering Bounds for Control Systems

s:fat

The VC-dimension characterizes learnability of $\{0, 1\}$-valued functions and by Sauer's Lemma it can be used to bound the growth function $S(\mathcal{A}, n)$. For learning real valued functions we look for a generalization of the VC-dimension with similar properties. In the previous section we calculated the pseudo-dimension that is related to learning real valued functions. Unfortunately, pseudo-dimension does not share the property the VC-dimension has; there are learnable function classes with infinite pseudo-dimension, see [39, p.206] and [8].

In this section we look for bounds on the fat-shattering dimension that corresponds to shattering with fixed "margin" γ. Both the pseudo-dimension and the fat-shattering dimension can be used to bound certain covering numbers and in this sense they act like the VC-dimension when bounding the growth function. Moreover, the fat-shattering dimension gives upper and lower bounds for covering numbers of function classes and the finiteness of the fat-shattering dimension can characterize learnability (see [4] and [5]).

We begin by proving a fat-shattering bound that is effective when the margin γ is "small" and apply it to derive fat-shattering bounds for continuous-time control systems discussed in the previous section.

2.2.1 A Fat-Shattering Bound

Definition 4. Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is *γ-shattered by $\mathcal{F}$* if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b_x indexed by X, there is a function $f_b \in \mathcal{F}$ satisfying

$$\begin{aligned} f_b(x) &\geq r_x + \gamma, && \text{if } b_x = 1, \text{ and} \\ f_b(x) &\leq r_x - \gamma, && \text{otherwise.} \end{aligned}$$

The *fat-shattering dimension* $\mathrm{fat}_\gamma(\mathcal{F})$ is a function from positive real numbers to integers which maps a value γ to the size of the largest fat-shattered set, if it is finite, or infinity otherwise.

We modify the above definition slightly to consider shattering at a fixed level.

Definition 5. Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is *level γ-shattered by $\mathcal{F}$* at level r if it can be γ-shattered when choosing the $r_x = r$ for all $x \in X$. The *level fat-shattering dimension* of $\mathcal{F}$ is defined analogously and it is denoted by $\mathrm{lfat}_\gamma(\mathcal{F})$.

If all functions in $\mathcal{F}$ have a bounded range $[a, b]$ then finiteness of $\mathrm{fat}_\gamma(\mathcal{F})$ and $\mathrm{lfat}_\gamma(\mathcal{F})$ hold simultaneously [4]. By modifying a proof by Gurvits [23] for range $[a, b]$ we see that

e:lfat

$$\left(2\frac{b-a-2\gamma}{\gamma}\right)^{-1} \mathrm{fat}_\gamma(\mathcal{F}) \leq \mathrm{lfat}_{\gamma/2}(\mathcal{F}) \leq \mathrm{fat}_{\gamma/2}(\mathcal{F}) \tag{2.2.14}$$

Now we illustrate how level fat-shattering dimension can be bounded for Lipschitz functions in certain cases:

t:lfat-gen

Theorem 2.2.1. *Let $F(\lambda, u) : \mathbb{R}^k \times U \to \mathbb{R}$ be such that $F(\cdot, u)$ is Lipschitz with constant L, i.e., $|F(\lambda_1, u) - F(\lambda_2, u)| \leq L\|\lambda_1 - \lambda_2\|$ for all $u \in U$, and consider the following class of functions*

$$\mathcal{F}_B = \{F(\lambda, \cdot) : U \to \mathbb{R}\ ;\ \lambda \in B\}.$$

Then

$$\begin{aligned}
\mathrm{lfat}_\gamma(\mathcal{F}_{B^k_\infty(C)}) &\leq k \log_2 \left\lfloor \frac{CL}{\gamma} \right\rfloor, \\
\mathrm{lfat}_\gamma(\mathcal{F}_{\bar{B}^k_\infty(C)}) &\leq k \log_2 \left(1 + \left\lfloor \frac{CL}{\gamma} \right\rfloor\right), \\
\mathrm{lfat}_\gamma(\mathcal{F}_{\bar{B}_{2,1}}) &\leq k \log_2 \left(C + \frac{L}{\gamma}\right),
\end{aligned}$$

where

$$\begin{aligned}
B^k_\infty(C) &= \{x \in \mathbb{R}^k\ ;\ \|x\|_\infty = \max_{1 \leq i \leq k} |x_i| < C\}, \\
\bar{B}^k_\infty(C) &= \{x \in \mathbb{R}^k\ ;\ \|x\|_\infty = \max_{1 \leq i \leq k} |x_i| \leq C\}, \\
\bar{B}^k_2(C) &= \{x \in \mathbb{R}^k\ ;\ \|x\|_2 = \sqrt{\textstyle\sum_{i=1}^k x_i^2} \leq C\}.
\end{aligned}$$

Proof. For the first part of the proof we use a generic set B for parameters. Assume that we can γ-shatter a set of inputs $\{u_1, \ldots, u_d\}$ at level r, i.e., for each assignment $b \in \{0, 1\}^d$ there exists a $\lambda \in B$ such that

$$\begin{aligned}
F(\lambda, u_i) &\geq r + \gamma, \quad &&\text{if } b_i = 1, \text{ and} \\
F(\lambda, u_i) &\leq r - \gamma, \quad &&\text{otherwise.}
\end{aligned}$$

We write $\lambda \sim \mu$ if and only if the parameters λ and μ give the same assignment for all $\{u_1, \ldots, u_d\}$. Further, let $\Lambda = \{\lambda_1, \ldots, \lambda_{2^d}\}$ be a collection of parameters that shatter $\{u_1, \ldots, u_d\}$ and let $\lambda_i, \lambda_j \in \Lambda$. Now $\lambda_i \not\sim \lambda_j$ implies that there exists $u^* \in \{u_1, \ldots, u_d\}$ such that $F(\lambda_i, u^*) \geq \gamma + r$ and $F(\lambda_j, u^*) \leq \gamma - r$, or vise versa. Hence

$$2\gamma \leq |F(\lambda_i, u^*) - F(\lambda_j, u^*)| \leq L\|\lambda_i - \lambda_j\|$$

and so $\|\lambda_i - \lambda_j\| \geq 2\gamma/L$. That is, the set Λ of candinality 2^d is a $2\gamma/L$-separated set in B. Now the level fat-shattering bounds follow by calculating $2\gamma/L$-packing numbers for the different sets B.

If $B = B^k_\infty(C)$ the maximum cardinality for ϵ-separated set is $\lfloor 2C/\epsilon \rfloor^k$ and thus

$$2^d \leq \left\lfloor \frac{2C}{2\gamma/L} \right\rfloor^k = \left\lfloor \frac{CL}{\gamma} \right\rfloor^k$$

and solving for d yields $d \leq k \log_2 \lfloor CL/\gamma \rfloor$.

Similarly if $B = \bar{B}^k_\infty(C)$ the maximum cardinality for ϵ-separated set is $(1 + \lfloor 2C/\epsilon \rfloor)^k$ and by similar argument we arrive to the bound $d \leq k \log_2 (1 + \lfloor LC/\gamma \rfloor)$.

For $B = \bar{B}^k_2(C)$ let $P(\epsilon)$ be a collection of ϵ-separated sets in $\bar{B}^k_2(C)$ and let $|P(\epsilon)|$ denote its cardinality. As all balls with radius $\epsilon/2$ with centers at ϵ-separated points have to be disjoint and their union have to be inside a ball of radius $C + \epsilon/2$, we get that $|P(\epsilon)|\alpha(k)(\epsilon/2)^k \leq \alpha(k)(C+\epsilon/2)^k$, where $\alpha(k) = \pi^{k/2}/\Gamma(k/2+1)$ is the volume of a unit ball in $\mathbb{R}^k$. Hence $|P(\epsilon)| \leq (C + 2/\epsilon)^k$ and

$$2^d \leq (C + L/\gamma)^k,$$

i.e., $d \leq k \log_2(C + L/\gamma)$. □

2.2.2 An Application to Control Systems

In this section we apply first the previous bounding technique to control systems for which we already have calculated the associated VC-dimension because for every $\gamma > 0$, the level fat-shattering dimension is also bounded by the VC-dimension. However, the above bounding technique does not require a special structure on input functions and we state a bound for L^1 controls too.

Let $U \subset \mathbb{R}^{mk}$ be a bounded set. Define a mapping $F : \lambda \times U \to [a, b]$ such that $F(\lambda, G) = y(\tau)$, where $\tau \geq 1$ and $y(\tau)$ is a solution of (2.2.6) with system parameters λ, initial condition $x(0) = 0$ and the control is given by G and the control functions ω_j are in Ω given by (2.2.2) and (2.2.3). Define a control system class by $\mathcal{F}_B = \{F(\lambda, \cdot) \colon U \to [a,b] \; ; \; \lambda \in B\}$ We assume further that A in (2.2.6) has some fixed Jordan block structure.

Theorem 2.2.2. *If for the above control system class is parametrized by* $B = \{\lambda \in \mathbb{R}^{n(m+1)} \; ; \; \|\lambda\|_\infty < 1\}$ *then*

$$\operatorname{lfat}_\gamma(\mathcal{F}_B) \leq \min \begin{cases} (m+1)n \log_2 \left\lfloor \frac{n^2 m \tau^{n+1} e^\tau k M}{\gamma} \right\rfloor, \\ 2(m+4)n \log_2 \Big(8e\big(nmk4(n+\ell_{\max}+1)+1\big)\big(3nk+2k+1\big)\Big), \end{cases}$$

where $\ell_{\max}$ *is given by* (2.2.2) *and* (2.2.3) *and*

$$\int_0^\tau |u_i(\tau - t)| dt \leq kM$$

for all $i = 1, \ldots, m$. *Moreover,*

$$\operatorname{fat}_\gamma(\mathcal{F}_B) \leq \frac{2}{\gamma}(b - a - 2\gamma) \operatorname{lfat}_{\gamma/2}(\mathcal{F}_B)$$

Proof. Our aim is to compute the Lipschitz constant associated to the above control system and then apply Theorem 2.2.1. For system $\dot{x} = Ax + Bu$, $y = Cx$ and initial condition $x(0) = 0$ we can write $Ce^{At}B = [\gamma_1, \ldots, \gamma_m]$, where each γ_i is a linear combination of n functions $\xi_1, \ldots, \xi_n$. Assume that A has a fixed Jordan block structure and let $a_k + ib_k$ to be an eigenvalue of A. We take $\alpha_{11}, \ldots, \alpha_{nm}, a_1, b_1, \ldots, a_r, b_r$ to be the system parameters, where $\gamma_i(t) = \sum_{j=1}^n \alpha_{ij}\xi_j(t)$ for $i = 1, \ldots, m$ and $a_1, b_1, \ldots, a_r, b_r$ are n eigenvalue parameters.

Denote the system parameters $(\alpha_{11}, \ldots, \alpha_{nm}, a_1, b_1, \ldots, a_r, b_r)$ by λ and assume $\|\lambda\|_\infty < 1$. Let $F(\lambda, u) = y(\tau) = \int_0^\tau \sum_{i=1}^m \sum_{\ell=1}^m \alpha_{i\ell}\xi_\ell(t) u_i(\tau - t) dt$. Functions $\xi_1(t), \ldots, \xi_n(t)$ are of the form $\xi(t) = t^c e^{at} \sin(bt)$ or $\xi(t) = t^c e^{at} \cos(bt)$ where $a + ib$ is an eigenvalue of A and $c \in \{0, \ldots, n\}$. Thus taking a partial derivative with respect to a or b will increase the power of t by one and change the trigonometric functions. Thus

$$\left|\frac{\partial F(\lambda,u)}{\partial \alpha_{\kappa\rho}}\right| = \left|\frac{\partial y(\tau)}{\partial \alpha_{\kappa\rho}}\right| = \left|\sum_{i=1}^{m}\sum_{\ell=1}^{n} d(i,\ell)\int_0^\tau \xi_\ell(t)u_i(\tau-t)dt\right|$$
$$\leq nm\int_0^\tau |\xi_\ell(t)u_i(\tau-t)|\,dt \leq nm\tau^n e^\tau kM,$$

because $\sup_{t\in[0,\tau]}|\xi_\ell(t)| \leq e^\tau\tau^n$ and $d(i,\ell) = \partial\alpha_{ij}/\partial\alpha_{\kappa\rho} = 1$ if $(i,\ell) = (\kappa,\rho)$ and zero otherwise. Similarly we calculate

$$\left|\frac{\partial F(\lambda,u)}{\partial a_\kappa}\right| = \left|\frac{\partial y(\tau)}{\partial a_\kappa}\right| \leq nm\tau^{n+1}e^\tau kM \text{ and}$$
$$\left|\frac{\partial F(\lambda,u)}{\partial b_\kappa}\right| = \left|\frac{\partial y(\tau)}{\partial b_\kappa}\right| \leq nm\tau^{n+1}e^\tau kM$$

as $\sup_{t\in[0,\tau]}\left|\frac{\partial\xi_\ell(t)}{\partial_\kappa}\right| \leq e^\tau\tau^{n+1}$ and $\sup_{t\in[0,\tau]}\left|\frac{\partial\xi_\ell(t)}{\partial b_\kappa}\right| \leq e^\tau\tau^{n+1}$.

Now the Lipschitz constant can be taken to be $L = n^2me^\tau\tau^{n+1}kM$ as

$$|F(\lambda,u) - F(\lambda^*,u)| = |\bigtriangledown F\cdot(\lambda-\lambda^*)| \leq L\|\lambda-\lambda^*\|_\infty.$$

The number of system parameters is at most $nm+n = (m+1)n$ and we get the level fat-shattering bound by applying Theorem 2.2.1 with space dimension $n(m+1)$ and $L = n^2me^\tau\tau^{n+1}kM$.

The remaining part of the proof comes from the VC-dimension bound. As we assumed that A has a fixed Jordan block structure, every entry of $e^{A(1-t)}$ is a linear combination of n functions $\xi_1(t),\ldots,\xi_n(t)$. (That is, we don't need to consider all possible functions over different Jordan block structures.) This implies that in the Goldberg-Jerrum argument of Section 2.1.2 we can take $\ell = mn+4n$, $d = nmk4(n+\ell_{\max})+1$ and $s = 3nk+2k+1$. Moreover, in that section the VC-dimension bounds were derived for time interval $[0,1]$. However, the upper bound follows from the number of system parameters and the degrees of polynomials to be evaluated. Changing the time interval to be $[0,\tau]$ means just that we replace the eigenvalue parameters (refering to the proof of Proposition 2.1.11) $a, b, e^a\cos b, e^a\sin b$ by $a\tau, b\tau, e^{a\tau}\cos b\tau, e^{a\tau}\sin b\tau$.

The last statement of the theorem follows from (2.2.14). $\square$

The structure of the input quaranteed that the VC-dimension was finite. Next we state a slight modification of the previous theorem but now the VC-dimension is not necessarily bounded.

Theorem 2.2.3. *Assume that the system $\dot{x} = Ax + Bu$, $y = Cx$, $x(0) = 0$ can be parametrized by $\lambda \in \mathbb{R}^{n(m+1)}$ and $\|\lambda\|_\infty < 1$. Let $F(\lambda, u) = y(\tau)$ be the solution with system parameters λ and control $u = (u_1, \ldots, u_m) \in U$. If each control u_i, $i = 1, \ldots, m$ satisfies*

$$\int_0^\tau |u_i(\tau - t)| \, dt \le M$$

and

$$\mathcal{F}_B = \{F(\lambda, \cdot) \colon U \to [a, b] \, ; \, \lambda \in B^k_\infty(C)\}$$

then

$$\operatorname{lfat}_\gamma(\mathcal{F}_B) \le n(m+1) \log_2 \left\lfloor \frac{n^2 m \tau^{n+1} e^\tau M}{\gamma} \right\rfloor$$

and

$$\operatorname{lfat}_\gamma(\mathcal{F}_B) \le \frac{2}{\gamma}(b - a - 2\gamma) n(m+1) \log_2 \left\lfloor \frac{2n^2 m \tau^{n+1} e^\tau M}{\gamma} \right\rfloor$$

Proof. We introduce system parameters and functions $\xi_1(t), \ldots, \xi_n(t)$ as in the previous proof. The only change is that we bound

$$\begin{aligned}
\int_0^\tau |\xi_l(t) u_i(\tau - t)| \, dt &\le \tau^n e^\tau M, \\
\int_0^\tau \left| \frac{\partial \xi_l(t)}{\partial a_\kappa} u_i(\tau - t) \right| \, dt &\le \tau^{n+1} e^\tau M, \\
\int_0^\tau \left| \frac{\partial \xi_l(t)}{\partial b_\kappa} u_i(\tau - t) \right| \, dt &\le \tau^{n+1} e^\tau M.
\end{aligned}$$

The fat shattering bound follows from (2.2.14) □

Remark 2.2.4. In a special case of scalar controls (i.e., $m = 1$) that are given by a linear combination of k fixed input functions, the control system discussed can be seen as a hyperplane in $\mathbb{R}^k$. Let $\mathcal{F} = \{x \in \mathbb{R}^k \mapsto \sum_i w_i x_i + \theta \ ; \ \|w\|_2 = 1\}$. If we further restrict x such that $\|x\|_2 \leq R$ and $|\theta| \leq R$ then [37, 34]

$$\mathrm{fat}_\gamma(\mathcal{F}) \leq \min\{9R^2/\gamma^2, k+1\} + 1.$$

In this special case our γ-shattering result is of the form $(c_1 + c_2/\gamma)\log_2(\lfloor c_3 k/\gamma \rfloor)$, where c_1, c_2 and c_3 are constants. This gives improvement over the hyperplane bound when the margin γ is small.

2.2.3 Structural Risk Minimization with Fat-Shattering Bounds

As for the VC-dimension our fat-shattering bound for the control system depends on the system dimension. Fat-shattering can be connected to classification in structural risk minimization as we did for concept learning. Consider a sequence of function classes $\mathcal{F}_{n_1} \subset \mathcal{F}_{n_2} \subset \cdots \subset \mathcal{F}_{n_d}$ corresponding to control systems in such a way that the system dimension increases in the sequence while the other structure remains constant.

Assume further that we know positive numbers p_{n_i} such that $\sum p_{n_i} = 1$ where p_{n_i} represents our prior estimate for the probability that the smallest class containing a consistent hypothesis is $\mathcal{F}_{n_i}$.

Let T_θ denote the threshold function at θ: $T_\theta \colon \mathbb{R} \to \{0,1\}$, $T_\theta(\alpha) = 1$ if and only if $\alpha > \theta$. For a class of functions $\mathcal{F}$, $T_\theta(\mathcal{F}) = \{T_\theta(f) \colon f \in \mathcal{F}\}$.

Theorem 2.2.5 t:afat-bd ([34]). *Consider a real valued function class $\mathcal{F}$ having fat-shattering function bounded above by the function* afat: $\mathbb{R} \to \mathbb{N}$ *which is continuous from the right. Fix $\theta \in \mathbb{R}$. If a learner correctly classifies m independently generated examples with $h = T_\theta(f) \in T_\theta(\mathcal{F})$ and $\gamma = \min|f(x_i) - \theta|$, then with confidence $1-\delta$ the expected error of h is bounded from above by*

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k \log_2\left(\frac{8em}{k}\right)\log_2(32m) + \log_2\left(\frac{8m}{\delta}\right)\right),$$

where $k = \mathrm{afat}_{\gamma/8}(\mathcal{F})$.

Corollary 2.2.6. *Accuracy ϵ and confidence δ can be obtained if the sample size is at least*

$$\max\left\{\frac{16k}{\epsilon}\log_2\left(\frac{128\sqrt{ke}}{\epsilon}\right), \frac{8}{\epsilon}\log_2\left(\frac{32}{\epsilon\delta}\right)\right\}.$$

Proof. To obtain the sample complexity for fat-shattering we write $\epsilon(m,k,\delta) \le \epsilon$, i.e.,

$$\frac{2}{m}\left(k\log_2\left(\frac{8em}{k}\right)\log_2(32m) + \log_2\left(\frac{8m}{\delta}\right)\right) \le \epsilon.$$

The above is satisfied if

$$\frac{2}{m}k\log_2\left(\frac{256em^2}{k}\right) \le \frac{\epsilon}{2}$$

and

$$\frac{2}{m}\log_2\left(\frac{8m}{\delta}\right) \le \frac{\epsilon}{2}.$$

We start by studying the first term and we write it as

$$\frac{4}{m}k\log_2\left(\frac{16\sqrt{e}m}{\sqrt{k}}\right) \le \frac{\epsilon}{2},$$

or

e:fat-samp1
$$m \ge \frac{8k}{\epsilon}\log_2\left(\frac{16\sqrt{e}m}{\sqrt{k}}\right). \tag{2.2.15}$$

Lemma 2.2.7 ([39, p. 81]). *Suppose $\alpha, \beta > 0$, $\alpha\beta > 2$, $m \ge 1$. Then $m \ge 2\alpha\log_2(\alpha\beta)$ implies $m > \alpha\log_2(\beta m)$.*

By using the above lemma with $\alpha = 8k/\epsilon$ and $\beta = 16\sqrt{e}/\sqrt{k}$ we see that (2.2.15) is satisfied if

$$m \ge \frac{16k}{\epsilon}\log_2\left(\frac{128\sqrt{ke}}{\epsilon}\right).$$

Similarly for the second term we write

$$m \geq \frac{4}{\epsilon} \log_2 \left(\frac{8m}{\delta} \right),$$

which is satisfied if

$$m \geq \frac{8}{\epsilon} \log_2 \left(\frac{32}{\epsilon\delta} \right).$$

Thus we require that

$$m \geq \max \left\{ \frac{16k}{\epsilon} \log_2 \left(\frac{128\sqrt{ke}}{\epsilon} \right), \frac{8}{\epsilon} \log_2 \left(\frac{32}{\epsilon\delta} \right) \right\}.$$

□

We modify Theorem 2.2.5 for the sequence of hypothesis classes as follows:

Theorem 2.2.8. *Let $\mathcal{F}_{n_i}$, $i = 1, \dots, d$ be a sequence of function classes with fat-shattering function bounded above by the function $\mathrm{afat}_\gamma(\mathcal{F}_{n_i})\colon \mathbb{R} \to \mathbb{N}$ which is continuous from the right. Fix $\theta \in \mathbb{R}$ and let p_{n_i} be positive numbers satisfying $\sum p_{n_i} = 1$. With probability $1-\delta$ over m independent samples for any n_i for which learner finds consistent hypothesis $h = T_\theta(f_{n_i}) \in T_\theta(\mathcal{F}_{n_i})$ and $\gamma = \min |f_{n_i}(x_i) - \theta|$, the generalization error of h is bounded above by*

$$\epsilon(m, k_{n_i}, p_{n_i}, \delta) = \frac{2}{m} \left(k_{n_i} \log_2 \left(\frac{8em}{k_{n_i}} \right) \log_2(32m) + \log_2 \left(\frac{8m}{p_{n_i}\delta} \right) \right),$$

where $k_{n_i} = \mathrm{afat}_{\gamma/8}(\mathcal{F}_{n_i})$.

Proof. Denote the generalization error of h by $\mathrm{er}_P(h) = P(h \Delta T)$, where T is an unknown target. We want to find $\tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta)$ so that

$$Pr\Big\{ x\ ;\ \exists n_i \exists h = T_\theta(f_{n_i}) \in T_\theta(\mathcal{F}_{n_i}) \text{ s.t. } h \text{ is consistent,}$$

$$\gamma_{n_i} = \min |f_{n_i}(x_i) - \theta|, \text{ and } \mathrm{er}_P(h) \geq \tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta) \Big\} \leq \delta.$$

By using the union bound the left-hand side of the above inequality is bounded by

$$\sum_{n_i} Pr\Big\{x\ ;\ \exists h = T_\theta(f_{n_i}) \in T_\theta(\mathcal{F}_{n_i}) \text{ s.t. } h \text{ is consistent,}$$
$$\gamma_{n_i} = \min|f_{n_i}(x_i) - \theta|\,,\ \text{and}\ \mathrm{er}_P(h) \geq \tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta)\Big\}.$$

By taking $\tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta) = \epsilon(m, k_{n_i}, p_{n_i}\delta)$ and applying Theorem 2.2.5 the probability on the right-hand side is at most $p_{n_i}\delta$. Hence

$$Pr\Big\{x\ ;\ \exists n_i \exists h = T_\theta(f_{n_i}) \in T_\theta(\mathcal{F}_{n_i}) \text{ s.t. } h \text{ is consistent,}$$
$$\gamma_{n_i} = \min|f_{n_i}(x_i) - \theta|\,,\ \text{and}\ \mathrm{er}_P(h) \geq \tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta)\Big\} \leq \sum_{n_i} p_{n_i}\delta \leq \delta,$$

with

$$\tilde{\epsilon}(m, k_{n_i}, p_{n_i}, \delta) = \epsilon(m, k_{n_i}, p_{n_i}\delta) = \frac{2}{m}\left(k_{n_i}\log_2\left(\frac{8em}{k_{n_i}}\right)\log_2(32m) + \log_2\left(\frac{8m}{p_{n_i}\delta}\right)\right).$$

□

Chapter 3

Learning an Open Interval with Side Information

c:si-int

In this chapter we introduce a new type of a learning setting motivated by a control application. In this problem a random sample generates new additional information, called side information, that may speed learning. We begin by defining the problem and discussing how it modifies the standard learning setting. In Section 3.2 we analyze a simple learning problem. It allows us to calculate exact learning rates instead of upper bounds and it also points to phenomena that hold more generally. Then Section 3.3 discusses asymptotic log learning rates and generalizes from the results obtained in the simple learning problem. The advantage of side information is highly distribution dependent and we end the chapter by illustrating the small sample effect of side information by numerical simulations with different distributions.

3.1 Learning with Side Information

Probably Approximately Correct (PAC) learning theory studies learning algorithms that train on randomly generated data sets with the goal of learning to classify new, randomly presented instances with a small probability of error. Standard PAC theory, as in [39], assumes that the training sample is a string of independent and identically distributed (i.i.d.) random variables and that the training data and the instances to be classified both come from the same probability distribution.

In this chapter we analyze a simple learning problem in which each training observation gives rise to a piece of additional information, called side information. However, the output of the algorithm will need only to classify a typical observation and not the side information. The side information problem is thus an example in which the training and the classification distributions differ. The learning algorithm will train by picking a concept among those consistent with a training multi-sample. When the side information is available, the algorithm chooses from a smaller and more probably accurate set of concepts, and hence should operate more efficiently. Our aim in this chapter is to study a simple model example to explore when and by how much side information can speed learning.

Our motivation to introduce this type of problem comes from a control application. Let us study a discrete-time time-invariant control system $x^+ = Ax + Bu$, $y = Cx$, where A, B and C are $n \times n$, $n \times 1$ and $1 \times n$ dimensional matrices, respectively. Assume further that the initial condition is $x_0 = 0$. Given an input $u = (u_1, \dots, u_k)$ we are interested in the final state $y_k = \sum_{i=1}^k CA^{i-1}Bu_i = \sum_{i=1}^k c_i u_i$. (Note that $\sum_{i=1}^k c_i u_i$ is an n-recursive perceptron for which DasGupta and Sontag have calculated VC-dimension upper and lower bounds [14].)

Our aim is to learn the mapping from inputs u to the finals states y_k. However, many times the intermediate states $y_1, \dots, y_{k-1}$ are also available and knowledge on the intermediate states may speed the learning of the final state mapping. Indeed, due to the initial state 0 the input $u = (0, u_1, \ldots, u_{k-1})$ yields the final state y_{k-1}, and continuing in the similar fashion $u = (0, \ldots, 0, u_1)$ gives rise to y_1. With the actual input-output behavior one noiseless sequence of generated inputs and outputs is enough to identify the system. With sign observations our aim is to learn the mapping $u = (u_1, \ldots, u_k) \mapsto \mathrm{sign}(y_k) = \mathrm{sign}(\sum_{i=1}^{k} CA^{i-1}Bu_i)$ and the side information consists of mappings $(0, u_1, \ldots, u_{k-1}) \mapsto \mathrm{sign}(y_{k-1}), \ldots, (0, \ldots, 0, u_1) \mapsto \mathrm{sign}(y_1)$.

Now we define the side information learning problem formally. We first recall the standard PAC learning model, using the notation of [39]. Let $\mathcal{C}$ be a concept class, a class of subsets of a space X, and let $T \in \mathcal{C}$ be an unknown target concept to be learned. An i.i.d. sample $x_1, \ldots, x_m$ is drawn according to a probability distribution P (known or unknown) on X and an oracle returns $I_T(x_1), \ldots, I_T(x_m)$, where I_T is the characteristic function of T. Based on the multi-sample $\{(x_1, I_T(x_1)), \ldots, (x_m, I_T(x_m))\}$ an algorithm forms an estimate h_m for the unknown T. Let the learning error be $\mathrm{Err}(T, h_m, m, \epsilon) = P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\}$, where $d_P(T, h_m) = P(T \Delta h_m)$ is the probability that h_m will misclassify a future random sample and P^m is the m-fold product of P. The quality of h_m is evaluated by the learning rate $lr(m, \epsilon) = \sup_{T \in C} \mathrm{Err}(T, h_m, m, \epsilon)$.

We modify the standard setup as follows: let $\bar{Z}_1, \ldots, \bar{Z}_m$ be an i.i.d. sequence such that $\bar{Z}_i = (X_i, Y_i) \in X \times Y$. The $\{X_i\}$ sequence is referred to as the original data, while the $\{Y_i\}$ is the sequence of side information. $\bar{z}_1, \ldots, \bar{z}_m$ is observed and a learning algorithm forms an estimate $\widetilde{h}_m$ for the target based on the multi-sample $\{(\bar{z}_1, I_T(\bar{z}_1)), \ldots, (\bar{z}_m, I_T(\bar{z}_m))\}$. The estimate $\widetilde{h}_m(\{(\bar{z}_1, I_T(\bar{z}_1)), \ldots, (\bar{z}_m, I_T(\bar{z}_m))\})$ is a subset of X and its quality is evaluated on how well it classifies in the X space. Let P_X and P_Z denote probabilities in X and Z spaces. We introduce the learning rate with side information as $\widetilde{lr}(m, \epsilon) = \sup_{T \in \mathcal{C}} P_Z^m\{\bar{z} \in Z^m \ ; \ d_{P_X}(T, \widetilde{h}_m) > \epsilon\}$.

Learning with side information has difficult features that can not be overemphasized:

- Side information gives rise to non-i.i.d. observations.
- The training space is different from the evaluation space.

Learning with non-i.i.d. data has been attempted only recently and different training and evaluation spaces break the beauty of distribution independent learning, if we want to show that side information speeds learning. In the standard learning setting if the probability of obtaining samples lying in some region of X is low, the estimate is likely to be poor in the same region. However, the probability that an unseen future sample falls in that region is also low and it does not affect the goodness of the estimate. With side information problems of "apples versus oranges" are inevitable—side information samples may be very different from the original samples. For example, assume that the learning algorithm operates only on positive samples. Then, depending on the distribution of the samples and the unknown target, it is possible that all side information samples are actually negative samples.

Next we shortly review papers that are closest to some aspects of the side information problem. In presence of side information the training phase utilizes non-i.i.d. data. Meir [29] has studied a learning problem in the context of nonlinear time series prediction with dependent data. He assumes data to satisfy a certain mixing condition and by a blocking scheme, see [42], the dependent sequence is mapped to the independent one. Due to the dependency in the data learning is slower than with an i.i.d. sample. In our case side information provides a natural blocking and our aim is to show that side information speeds learning. For dependent data see also the paper by Campi and Kumar [13].

Also one can consider training two algorithms: one with original data and one with side information. This relates to the work by Blum and Mitchell [12] in which there are two views (x_1, x_2) of an example. A necessary assumption in the training is that x_1 and x_2 are not completely correlated. This excludes the side information setting, where the natural pairing would be $(x, s(x))$. Abu-Mostafa [1, 2, 3] has studied learning with hints, where hints are properties of the target function known to us. He discusses different types of hints; virtual examples and duplicate examples, and various methods to incorporate hints into practical learning applications. He shows how hints can speed learning in a financial market application. Ratsaby and Maiorov [31] have extended PAC learning theoretical analysis to problems in which there is some form of *a priori* partial information about the target available in addition to the random sample. The learning algorithm is first shown n information operations and then m labeled samples. This work connects approximation theory to the PAC learning.

3.2 Learning an Open Interval with Side Information

s:int

As a preliminary investigation into the effect of side information we explore a very basic example with a nice geometrical structure. We consider an open target interval $T = (a, b) \subset (0, 1)$ and a training scheme in which each X_i drawn from $(0, 1)$ give rise to $s(X_i)$, where

$$s(x) = (x + \frac{1}{2}) \mod 1.$$

The oracle returns not only $I_T(x_i)$, but also $I_T(s(x_i))$ for $i = 1, \ldots, m$. We study a fixed algorithm where the estimate for the target is the smallest interval containing all positive samples, see Figure 3.1. Later in this chapter we consider a more general learning algorithm when we study the asymptotic advantage of side information. Further, in Chapter 4 our aim is to generalize from the results obtained in this chapter.

When we estimate the target by the smallest interval containing positive samples, the estimate with side information can never be worse than the one without it. For

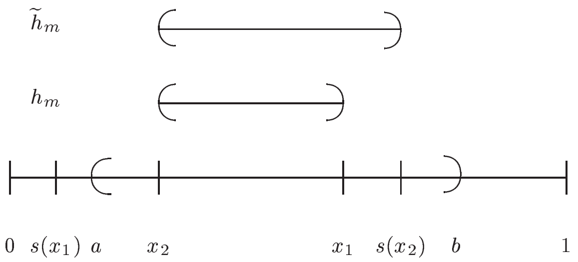

Figure 3.1: Estimating the target interval $T = (a, b)$: h_m uses the original i.i.d. sample and $\widetilde{h}_m$ takes advantage of the side information.

f:alg

uniform distribution the side information mapping is measure preserving and the intuition suggest that the learning rate should correspond to having twice as large a sample. Unfortunately this intuition ignores the fact that samples are highly correlated and we will discover a relation between the side information mapping and the target concept. Our purpose is to calculate exact learning rates with and without side information and compare to quantify the improvement of the learning rate and how it relates to the problem structure.

First we establish the learning rate $lr(m, \epsilon)$ in the standard setup. The calculation is repeated for the uniform distribution when the side information is used together with the original sample. Somewhat heavy machinery is used in calculating the learning rate in the standard setting. However, that enables us to handle the more complicated case of side information fairly easily. Comparing learning rates shows that adding side information speeds considerably the learning in the case of uniform distribution.

Translation by 1/2 was chosen because of simplicity and naturality: when learning an interval on a circle, translation by 1/2 corresponds to adding the antipodal point as side information. The interesting phenomena, namely the the dependencies of learning error improvement on both the concept and distribution, occur already in this special case. Arbitrary translations mod 1 can be mostly reduced to problems studied here. The only new phenomenon is the non-symmetric overlap of the target interval and the translated interval. Still, a result on learning errors similar to one demonstrated in this work (asymptotic exponential rate) holds. Sections 3.2.1 and 3.2.2 present the exact rates and Section 3.2.3 provides the comparison and conclusions.

3.2.1 Learning Rates without Side Information

s:int-nosi

We begin by establishing the learning rate for the interval learning problem without side information.

Fix $\epsilon > 0$ and let $X_1, \ldots, X_m$ be an i.i.d. random variables with probability density $g(x)$, $g(x) > 0$ on (0,1). Let $T = (a, b)$ such that $\int_a^b g(s)ds > \epsilon$ as otherwise the trivial estimate $h_m = \emptyset$ would be enough. We calculate the learning rate $lr(m, \epsilon)$ by first showing that it does not depend on the distribution P_g. After that we study the uniform distribution on (0,1) and we show through a series of small propositions that it is enough to consider the target $T = (0, 1)$. Keeping in mind that the learning algorithms operates on positive samples the fact that the learning rate does not depend on the length of the target interval may sound unexpected. However, in learning the estimate of the target is either "bad" (i.e., the misclassification probability is greater than ϵ) or "good". In order to obtain a "good" estimate, samples have to fall close to the endpoints of the target interval. As the problem has been reduced to studying the uniform distribution, the location of the endpoints is irrelevant.

The main result of this section is:

t:noside

Theorem 3.2.1. *Let P_g have a density $g(x)$ with $g(x) > 0$ on $(0, 1)$. Then for all targets $T = (a, b)$, with $\int_a^b g(s)ds > \epsilon$,*

$$P_g^m\{\bar{x} \in X^m : d_{P_g}(T, h_m) > \epsilon\} = (1-\epsilon)^m + m\,\epsilon(1-\epsilon)^{m-1}.$$

and hence

$$lr(m, \epsilon) = \sup_{T \in \mathcal{C}} P_g^m\{\bar{x} \in X^m : d_{P_g}(T, h_m) > \epsilon\} = (1-\epsilon)^m + m\,\epsilon(1-\epsilon)^{m-1}.$$

Proof. The theorem follows from Lemma 3.2.3 and Propositions 3.2.2, 3.2.5 and 3.2.6 that are stated and proved next. □

By considering purely atomic measures P, we observe that $d_P(T, h_m)$ is a discrete random variable and the statement of Theorem 3.2.1 can not be generalized for such measures.

p:re-to-u

Proposition 3.2.2. *Let P_g have a density $g(x)$ with $g(x) > 0$ on $(0, 1)$. Then for all targets $T = (a, b)$, $\int_a^b g(s)ds > \epsilon$,*

$$P_g^m\{\bar{x} \in X^m : d_{P_g}(T, h_m) > \epsilon\} = P^m\{\bar{x} \in X^m \ ; \ d_P((f(a), f(b)), h_m) > \epsilon\},$$

where P stands for the uniform distribution on $(0, 1)$ and $f(x) = \int_0^x g(s)ds$.

Proof. Apply the transformation $f(x) = \int_0^x g(s)ds$ to $x_1, \ldots, x_m$ that are drawn according to $g(x)$. Then $f(x_1), \ldots, f(x_m)$ are from the uniform distribution and

$$\begin{aligned} P_g^m\{\bar{x} \in X^m : d_{P_g}(T, h_m) > \epsilon\} & \\ &= P^m\{f(\bar{x}) \in X^m : d_P((f(a), f(b)), h_m(f(\bar{x}))) > \epsilon\} \\ &= P^m\{\bar{x} \in X^m \ ; \ d_P((f(a), f(b)), h_m) > \epsilon\}. \end{aligned}$$

□

From now on, in this chapter P refers to the uniform distribution and P^m to its m-fold product. The next lemma states what the learning rate is.

l:lrate

Lemma 3.2.3. $P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\} = (1-\epsilon)^m + m\epsilon(1-\epsilon)^{m-1}$ *for all target intervals $T = (a, b)$ with $b - a > \epsilon$, and hence* $\sup_{T \in \mathcal{C}} P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\} = (1-\epsilon)^m + m\epsilon(1-\epsilon)^{m-1}$.

Proof. First, it is enough to consider the target interval $T = (0,1)$: Propositions 3.2.5 and 3.2.6 (that are stated and proved later) show that $P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\}$ is independent of the target T. When m samples on $(0,1)$ are drawn independently from the uniform distribution and ordered, we get the order statistics under the uniform distribution. Let $X_{i;m}$ denote the ith order statistic from the sample of size m. By using that the range of the (0,1) uniform order statistics, $X_{m;m} - X_{1;m}$, has the density $m(m-1)x^{m-2}(1-x)$, $0 \leq x \leq 1$, see [7],

$$\begin{aligned} P^m\{\bar{x} \in & X^m \ ; \ d_P(T, h_m) > \epsilon\} \\ &= Pr\{X_{m;m} - X_{1;m} < 1 - \epsilon\} \\ &= \int_0^{1-\epsilon} m(m-1)\tau^{m-2}(1-\tau)d\tau \\ &= (1-\epsilon)^m + m\epsilon(1-\epsilon)^{m-1}. \end{aligned}$$

□

Remark 3.2.4. Consider the target interval $T = (0, b)$ and let $b \searrow \epsilon$. Then with probability one $d_P(T, h_m) > \epsilon$ if and only if an estimate for h_m has not been formed, i.e., less than two samples fall in $(0, b)$. Then

$$\lim_{b \searrow \epsilon} \Pr\{\text{less than 2 samples fall in (0,b)}\} = (1-\epsilon)^m + m\epsilon(1-\epsilon)^{m-1}.$$

The first term in the sum corresponds to the probability that no sample falls in $(0,\epsilon)$ and the second term gives the probability that exactly one sample falls in $(0, \epsilon)$.

We are left to justify that $P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\}$ does not depend on the target T. Let $T = (a, b)$, with $b - a > \epsilon$. First we show that the interval can be moved to begin at 0 and then we show that $\mathrm{Err}(T, h_m, m, \epsilon) = P^m\{\bar{x} \in X^m \ ; \ d_P((0, b), h_m) > \epsilon\}$ is independent of b. The following sequence of transformations and swapping of intervals is later used again in studying learning with side information.

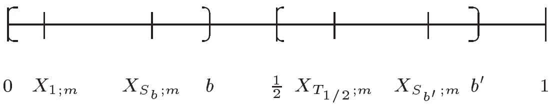

Figure 3.2: The stopping time notation used in proofs.

f:not

p:int0b

Proposition 3.2.5. *For $b - a > \epsilon$*

$$P^m\{\bar{x} \in X^m \ ; \ d_P((a,b), h_m(\bar{x})) > \epsilon\} = P^m\{\bar{x} \in X^m \ ; \ d_P((0, b-a), h_m(\bar{x})) > \epsilon\}.$$

Proof. Let f be the map defined by $f(x) = x - a$, if $x \geq a$ and $f(x) = x + 1 - a$ otherwise. f is a bijection that maps the interval (a, b) to $(0, b-a)$ and it preserves the uniform distribution on (0,1). Moreover $f(x_1), f(x_2), \ldots f(x_m)$ are i.i.d. on (0,1). Hence

$$\begin{aligned} P^m\{\bar{x} \in & X^m \ ; \ d_P((a,b), h_m(\bar{x})) > \epsilon\} \\ &= P^m\{f(\bar{x}) \in X^m \ ; \ d_P((0, b-a), h_m(f(\bar{x}))) > \epsilon\} \\ &= P^m\{\bar{x} \in X^m \ ; \ d_P((0, b-a), h_m(\bar{x})) > \epsilon\}. \end{aligned}$$

□

For $X_{1;m}$, $X_{2;m}$, ..., $X_{m;m}$, the order statistics on (0,1) under the uniform distribution, define a stopping time $T_z = \inf_k\{X_{k;m} > z\}$, $T_1 = m+1$, and let $S_z = T_z - 1$, see Figure 3.2. For notational purposes let us define $X_{0;m} = 0$, so that S_z is well defined even when $T_z = 1$. Note that $X_{S_b;m}$ is the closest sample on the left of b and we can write

$$P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), h_m(\bar{x})) > \epsilon\} \Leftrightarrow Pr\{X_{1;m} + (b - X_{S_b;m}) > \epsilon\}.$$

p:intlength

Proposition 3.2.6. *$P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), h_m(\bar{x})) > \epsilon\}$ does not depend on b for $b > \epsilon$.*

Proof. To prove the proposition, the argument of Equation (3.3.1) and the reasoning after it is enough. However, a swapping argument is presented, and later when a similar proposition is stated in the case of side information the swapping technique can be applied.

Without loss of generality assume that $\widehat{b} > b$. First assume that $b \geq 2\epsilon$ and fix $\widehat{b} > 2\epsilon$. Let f be a bijective Borel measurable map that switches intervals $(b-\epsilon, b)$ and $(\widehat{b}-\epsilon, \widehat{b})$. For simplicity assume that the intervals $(b-\epsilon, b)$ and $(\widehat{b}-\epsilon, \widehat{b})$ do not overlap. $f(x_1), f(x_2), \ldots, f(x_m)$ is an i.i.d. sequence on (0,1) and let $Y_{1;m}, Y_{2;m}, \ldots, Y_{m;m}$ denote its order statistics. Now

$$\begin{aligned}
P^m\{\bar{x} \in & X^m \ ; \ d_P((0,b), h_m(\bar{x})) > \epsilon\} \\
&= Pr\{X_{1;m} + (b - X_{S_b;m}) > \epsilon\} \\
&= Pr\{Y_{1;m} + (\widehat{b} - Y_{S_{\widehat{b}};m}) > \epsilon\} \\
&= P^m\{f(\bar{x}) \in X^m \ ; \ d_P((0,\widehat{b}), h_m(f(\bar{x}))) > \epsilon\} \\
&= P^m\{\bar{x} \in X^m \ ; \ d_P((0,\widehat{b}), h_m(\bar{x})) > \epsilon\}.
\end{aligned}$$

The above argument has to be refined for the case $\epsilon < b < 2\epsilon$, as in this case it is possible that $Y_{1;m} + \widehat{b} - Y_{S_{\widehat{b}};m} < \epsilon$ even when $X_{1;m} + (b - X_{S_b;m}) > \epsilon$, and vice versa. If $X_{1;m} < b - \epsilon$ this can not happen and the proof given above applies to this case. Now assume that $b - \epsilon < X_{1;m} < b$.

Let us define, for $x_1 < \epsilon$,

$$g_b(x_1) = Pr\{b - \widetilde{X}^{x_1}_{S_b;m-1} > \epsilon - x_1\},$$

where $\widetilde{X}^{x_1}_{i;m-1}$ is the ith order statistic of a sample of size $m-1$ drawn from the uniform distribution on $(x_1, 1)$ and we define $\widetilde{X}^{x_1}_{0;m-1} = x_1$. We have

e:cond-on-x1

$$\begin{aligned}
P^m\{\bar{x} \in X^m \ ; \ & d_P((0,b), h_m(\bar{x})) > \epsilon\} \\
&= E[\mathbf{1}_{\{X_{1;m} \geq \epsilon\}}] + E[\mathbf{1}_{\{X_{1;m} < \epsilon\}}] E[\mathbf{1}_{\{X_{1;m} + (b - X_{S_b;m}) > \epsilon\}} | X_{1;m}]. \quad (3.3.1)
\end{aligned}$$

The order statistics $X_{2;m}, \ldots, X_{m;m}$ conditioned on $X_{1;m} = x_{1;m}$ is the same as the order statistics of a $m-1$ sample on $(x_{1;m}, 1)$ (see [7, p. 15]). Hence $E[\mathbf{1}_{\{X_{1;m}+(b-X_{S_b;m})>\epsilon\}}|X_{1;m}] = g_b(X_{1;m})$ and further,

$$g_b(x_1) = (1 - x_{1;m} - (\epsilon - x_{1;m}))^{m-1} = (1-\epsilon)^{m-1},$$

so it is independent of b.

To give an alternative proof by the swapping argument let $X_{1;m} = x_1$ and swap the interval (x_1, b) with $(\widehat{b} - b + x_1, \widehat{b})$. Call the order statistics after swapping $Y_{i;m}$ and note that $Y_{1;m} = X_{1;m}$. After the swapping the equation (3.3.1) can be written as

$$E[\mathbf{1}_{\{Y_{1;m}\geq\epsilon\}}] + E[\mathbf{1}_{\{Y_{1;m}<\epsilon\}}]\, g_{\widehat{b}}(Y_{1;m}) = P^m\{\bar{x} \in X^m \; ; \; d_P((0,\widehat{b}), h_m) > \epsilon\}.$$

Finally, at the beginning we assumed that $b < \widehat{b}$. If we wish to swap intervals so that $\widehat{b} < b$, we can apply the previous arguments by interchanging the role of b and $\widehat{b}$.

We also assumed that swapped intervals do not overlap. If this is not the case, we can take an intermediate b^* so that the intervals on the left-side of b and $\widehat{b}$ do not overlap with an interval on the left-side of b^*. Then the above arguments show that $P^m\{\bar{x} \in X^m \; ; \; d_P((0,b), h_m) > \epsilon\} = P^m\{\bar{x} \in X^m \; ; \; d_P((0,b^*), h_m) > \epsilon\}$ and similarly for intervals $(0,\widehat{b})$ and $(0,b^*)$. Hence $P^m\{\bar{x} \in X^m \; ; \; d_P((0,b), h_m) > \epsilon\} = P^m\{\bar{x} \in X^m \; ; \; d_P((0,\widehat{b}), h_m) > \epsilon\}$. □

3.2.2 Learning with Side Information Under the Uniform Distribution: Exact Results

s:int-si

In this section we restrict our analysis to the uniform distribution and we modify the problem of learning a target interval $T = (a, b)$ by adding side information.

That is, each sample x_i gives us $I_T(x_i)$ and also $I_T(s(x_i))$, where $s(x) = (x + 1/2) \mod 1$. The estimate for the target with side information $\widetilde{h}_m$ is the smallest interval containing all positive samples (original samples and side information samples).

We will show that $P^m\{\bar{x} \in X^m \ ; \ d_P((a,b), \widetilde{h}_m) > \epsilon\}$ does not depend on the location of the target interval but it does depend on its length. The following result shows the learning speed with the side information.

t:sideinfo

Theorem 3.2.7 (Summary of results). *Let $b - a = \ell$. For learning under the uniform distribution and side information $s(x) = (x + 1/2) \mod 1$,*

$$P^m\{\bar{x} \in X^m \ ; \ d_P((a,b), \widetilde{h}_m(\bar{x})) > \epsilon\}$$
$$= \begin{cases} (1-2\epsilon)^m + 2m\epsilon(1-2\epsilon)^{m-1}, & \text{if } \ell \in (\epsilon, \frac{1}{2}] \cup (\frac{1}{2}+\epsilon, 1], \\ 2^m(\ell-\epsilon)^{m-1}\left(\ell - \epsilon + m(1+\epsilon-2\ell)\right), & \text{if } \ell \in (\frac{1}{2}, \frac{1}{2}+\frac{\epsilon}{2}), \\ 2^m(1-\ell)^m, & \text{if } \ell \in [\frac{1}{2}+\frac{\epsilon}{2}, \frac{1}{2}+\epsilon]. \end{cases}$$

The proof of Theorem 3.2.7 is split into propositions and lemmata. Recall that we have data $x_1, \ldots, x_m$ and its side information translates $s(x_1), \ldots, s(x_m)$ and our aim is to learn the target interval $T = (a,b)$. In a sense, we have two sets of data and one interval. In order to deal with the non-i.i.d. data, we modify the problem in such a way that we have i.i.d. $x_1, \ldots, x_m$ as a data but we work with two intervals, $T = (a,b)$ and $T' = (s^{-1}(a), s^{-1}(b)) = (s(a), s(b)) = (a', b')$. T' is chosen so that when a sample falls in T' its side information translate falls inside the original target interval T. The estimate for the target T is based on how samples fall inside the intervals T and T'. Note that when $a' > b'$, (a', b') actually denotes $(a', 1) \cup (0, b')$.

By using the stopping times defined in the previous section we can state the event $d_P(T, \widetilde{h}_m) > \epsilon$ as follows:

$$d_p(T, \widetilde{h}_m) > \epsilon \Leftrightarrow (Z_{T_a;2m} - a) \wedge (Z_{T_{a^*};2m} - a^*) + (b - Z_{S_b;2m}) \wedge (b^* - Z_{S_{b^*};2m}) > \epsilon,$$

where $x \wedge y = \min\{x, y\}$, $a^* = a + 1/2$, $b^* = b + 1/2$ and $Z_{1;2m}, \ldots, Z_{2m;2m} = X_{1;m}, \ldots, X_{m;m}, X_{1;m} + 1, \ldots, X_{m;m} + 1$. The data has been extended periodically over 1 because of the calculus mod 1. For example, if a' is close to 1, the closest sample on the right of a' may be a sample close to 0 and the distance has to be adjusted to the mod 1 calculus.

As in learning without the side information we first show that the target interval can be moved to begin at 0.

Proposition 3.2.8. *For $b - a > \epsilon$,*

$$P^m\{\bar{x} \in X^m \ ; \ d_P((a, b), \widetilde{h}_m) > \epsilon\} = P^m\{\bar{x} \in X^m \ ; \ d_P((0, b - a), \widetilde{h}_m) > \epsilon\}.$$

Proof. Repeat the argument of Proposition 3.2.5. $\square$

Proposition 3.2.9. *The learning error $P^m\{\bar{x} \in X^m \ ; \ d_P((0, b), \widetilde{h}_m) > \epsilon\}$ is independent of b for $\epsilon < b \leq \frac{1}{2}$ or $\frac{1}{2} + \epsilon < b \leq 1$.*

Proof. For $\epsilon < b \leq \frac{1}{2}$, we have

$$\begin{aligned} P^m\{\bar{x} \in & X^m \ ; \ d_P((0, b), \widetilde{h}_m(\bar{x})) > \epsilon\} \\ &= \Pr\{X_{1;m} \wedge (X_{T_{1/2};m} - \frac{1}{2}) + (b - X_{S_b;m}) \wedge (b' - X_{S_{b'};m}) > \epsilon\} \\ &= P^m\{\bar{x} \in X^m \ ; \ d_P((0, b + \frac{1}{2}), \widetilde{h}_m(\bar{x})) > \epsilon\}, \end{aligned}$$

as $b' = (b + \frac{1}{2}) \mod 1 = b + \frac{1}{2}$ and $(b + \frac{1}{2})' = b$. Therefore it is enough to prove the proposition only for $\epsilon < b \leq \frac{1}{2}$.

Fix $\widehat{b} > b$ such that $\epsilon < \widehat{b} \leq \frac{1}{2}$. The aim is to swap $(b - \epsilon, b)$ with $(\widehat{b} - \epsilon, \widehat{b})$ and $(b + 1/2 - \epsilon, b + 1/2)$ with $(\widehat{b} + 1/2 - \epsilon, \widehat{b} + 1/2)$. This will leave 0 and $\frac{1}{2}$ fixed and data on $(0, \frac{1}{2})$ remains on the same interval; similarly for $(\frac{1}{2}, 1)$. Therefore we can apply the argument of Proposition 3.2.6 to swap the intervals. If necessary we condition on the value of $X_{1;m}$ or $X_{T_{1/2};m}$.

Assume first that $2\epsilon < b < 1/2$. Note that $b' = b + 1/2$ and $\widehat{b}' = \widehat{b} + 1/2$. Let f be the bijection that swaps $(b - \epsilon, b)$ with $(\widehat{b} - \epsilon, \widehat{b})$ and $(b' - \epsilon, b')$ with $(\widehat{b}' - \epsilon, \widehat{b}')$.

After applying f call the resulting order statistics $Y_{1;m}, \ldots, Y_{m;m}$. Now

$$\begin{aligned}
&P^m\{\bar{x} \in X^m \;;\; d_P((0,b), \widetilde{h}_m(\bar{x})) > \epsilon\} \\
&\quad = \Pr\{X_{1;m} \wedge (X_{T_{1/2};m} - \frac{1}{2}) + (b - X_{S_b;m}) \wedge (b' - X_{S_{b'};m}) > \epsilon\} \\
&\quad = \Pr\{Y_{1;m} \wedge (Y_{T_{1/2};m} - \frac{1}{2}) + (\widehat{b} - Y_{S_{\widehat{b}};m}) \wedge (\widehat{b}' - Y_{S_{\widehat{b}'};m}) > \epsilon\} \\
&\quad = \Pr\{f(\bar{x}) \in X^m \;;\; d_P((0,\widehat{b}), \widetilde{h}_m) > \epsilon\} \\
&\quad = P^m\{\bar{x} \in X^m \;;\; d_P((0,\widehat{b}), \widetilde{h}_m) > \epsilon\}.
\end{aligned}$$

If $\epsilon < b < 2\epsilon$ we can write

$$\begin{aligned}
&P^m\{\bar{x} \in X^m \;;\; d_P((0,b), \widetilde{h}_m) > \epsilon\} \\
&\quad = E\left[\mathbf{1}_{X_{1;m} \wedge (X_{T_{1/2};m} - 1/2) > \epsilon}\right] + E\left[\mathbf{1}_{X_{1;m} \wedge (X_{T_{1/2};m} - 1/2) \le \epsilon}\right] \\
&\qquad \times E\left[\mathbf{1}_{X_{1;m} \wedge (X_{T_{1/2};m} - 1/2) + (b - X_{S_b;m}) \wedge (b' - X_{S_{b'}m;}) > \epsilon} \mid X_{1;m}, X_{T_{1/2};m}\right].
\end{aligned}$$

Let $X_{1;m} = x_1$ and $X_{T_{1/2};m} = x_{1/2}$. We apply the swapping argument to swap $(g_1(x_1), b)$ with $(\widehat{b} - b - g(x_1), \widehat{b})$ and $(g_2(x_{1/2}), b')$ with $(\widehat{b}' - b' - g_2(x_{1/2}), \widehat{b}')$, where $g_1(x_1) = x_1$, if $x_1 \in (b - \epsilon, b)$ and $b - \epsilon$ otherwise and $g_2(x_{1/2}) = x_{1/2}$, if $x_{1/2} \in (b' - \epsilon, b')$, and $b' - \epsilon$ otherwise. Denote the swapping function by f and let $Y_{1;m}, \ldots, Y_{m;m}$ be the resulting order statistics of the swapped sequence. Then

$$\begin{aligned}
&P^m\{\bar{x} \in X^m \;;\; d_P((0,b), \widetilde{h}_m) > \epsilon\} \\
&\quad = E\left[\mathbf{1}_{Y_{1;m} \wedge (Y_{T_{1/2};m} - 1/2) > \epsilon}\right] + E\left[\mathbf{1}_{Y_{1;m} \wedge (Y_{T_{1/2};m} - 1/2) \le \epsilon}\right] \\
&\qquad \times E\left[\mathbf{1}_{Y_{1;m} \wedge (Y_{T_{1/2};m} - 1/2) + (\widehat{b} - Y_{S_{\widehat{b}};m}) \wedge (\widehat{b}' - Y_{S_{\widehat{b}'};m}) > \epsilon} \mid Y_{1;m}, Y_{T_{1/2};m}\right] \\
&\quad = \Pr\{f(\bar{x}) \in X^m \;;\; d_P((0,\widehat{b}), \widetilde{h}_m) > \epsilon\} = P^m\{\bar{x} \in X^m \;;\; d_P((0,\widehat{b}), \widetilde{h}_m) > \epsilon\}.
\end{aligned}$$

Hence we conclude that

$$P^m\{\bar{x} \in X^m \;;\; d_P((0,b), \widetilde{h}_m) > \epsilon\} = P^m\{\bar{x} \in X^m \;;\; d_P((0,\widehat{b}), \widetilde{h}_m) > \epsilon\}.$$

$\square$

l:2eps

Lemma 3.2.10. *For $\epsilon < b \leq \frac{1}{2}$ or $\frac{1}{2} + \epsilon < b \leq 1$,*

$$P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), \widetilde{h}_m) > \epsilon\} = (1-2\epsilon)^m + 2m\epsilon(1-2\epsilon)^{m-1}.$$

Proof. We have

$$\begin{aligned} P^m\{\bar{x} \in & X^m \ ; \ d_P((0,b), \widetilde{h}_m) > \epsilon\} \\ &= \lim_{b \searrow \epsilon} P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), \widetilde{h}_m) > \epsilon\} \\ &= Pr\{\text{less than 2 samples fall on } (0,\epsilon) \cup (\frac{1}{2}, \frac{1}{2} + \epsilon)\} \\ &= (1-2\epsilon)^m + 2m\epsilon(1-2\epsilon)^{m-1}, \end{aligned}$$

where the first term in the sum on the last line is the probability of having no samples in $(0,\epsilon) \cup (\frac{1}{2}, \frac{1}{2}+\epsilon)$ and the second is the probability of having exactly one sample in $(0,\epsilon) \cup (\frac{1}{2}, \frac{1}{2}+\epsilon)$. □

Lemma 3.2.11.

$$\begin{aligned} P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), \widetilde{h}_m(\bar{x})) > \epsilon\} \\ = \begin{cases} 2^m(b-\epsilon)^{m-1}(b-\epsilon+m(1+\epsilon-2b)), & \textit{if } \frac{1}{2} < b < \frac{1}{2} + \frac{\epsilon}{2}, \\ 2^m(1-b)^m, & \textit{if } \frac{1}{2} + \frac{\epsilon}{2} \leq b \leq \frac{1}{2} + \epsilon. \end{cases} \end{aligned} \tag{3.3.2}$$

Proof. Let $T = (0,b)$, with $\frac{1}{2} < b \leq \frac{1}{2} + \epsilon$. Note that any data in $T \cap (\frac{1}{2}, b')$ implies that $d_P(T, \widetilde{h}_m) \leq \epsilon$. For $d_P(T, \widetilde{h}_m) > \epsilon$ to hold, necessarily all m samples have to fall in $(b', \frac{1}{2}) \cup (b, 1)$. Apply the following bijection to eliminate the gap at $(\frac{1}{2}, b)$, see Figure 3.3:

$$f(x) = \begin{cases} x - (1-b), & \text{if } \frac{1}{2} < x < b, \\ x + b - \frac{1}{2}, & \text{if } b - \frac{1}{2} < x < \frac{1}{2}, \\ x, & \text{otherwise.} \end{cases}$$

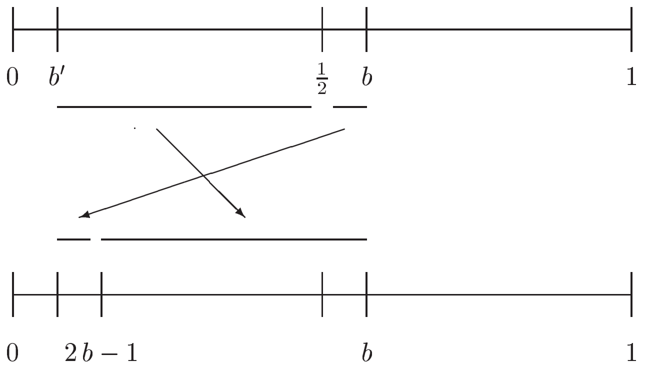

Figure 3.3: A measure preserving transformation to eliminate the gap at $(\frac{1}{2}, b)$.

f:gap

Let $Y_{i;m}$ refer to the order statistics after the transformation. The order statistics are from the uniform distribution on $(2b-1, 1)$ with midpoint b. The two eliminated intervals of length $b - \frac{1}{2}$ have to be taken into account when calculating error:

e:b12

$$\begin{aligned} P^m\{\bar{x} \in X^m \; ; \; d_P(T, \widetilde{h}_m) > \epsilon\} \\ = Pr\{(Y_{1;m} - (2b-1)) \wedge (Y_{T_b;m} - b) \\ + (b - Y_{S_b;m}) \wedge (1 - Y_{m;m}) > \epsilon - (2b-1) \\ \cap \text{ no samples in } (0, 2b-1)\} \end{aligned} \tag{3.3.3}$$

By using that $Y_{i;m} = (2b-1) + 2(1-b)X_{i;m}$, where $X_{i;m}$ are the order statistics on $(0, 1)$, equation (3.3.3) is equal to

$$Pr\left\{X_{1;m} \wedge (X_{S_{1/2};m} - \frac{1}{2}) + (\frac{1}{2} - X_{T_{1/2};m}) \wedge (1 - X_{m;m}) > \frac{\epsilon - (2b-1)}{2(1-b)}\right\} \times (2(1-b))^m.$$

If $1/2 + \epsilon/2 \leq b \leq 1/2 + \epsilon$, $\epsilon - (2b-1) \leq 0$, the first term equals 1 and $P^m\{\bar{x} \in X^m \; ; \; d_P(T, \widetilde{h}_m) > \epsilon\} = 2^m(1-b)^m$. For $\frac{1}{2} < b < \frac{1}{2} + \frac{\epsilon}{2}$ the first term can be obtained from Lemma 3.2.10 by setting $b = 1/2$ and $\epsilon' = \frac{\epsilon-(2b-1)}{2(1-b)}$. After substitution and simplification we get

$$P^m\{\bar{x} \in X^m \; ; \; d_P(T, \widetilde{h}_m(\bar{x})) > \epsilon\} = 2^m(b-\epsilon)^{m-1}(b - \epsilon + m(1 + \epsilon - 2b)).$$

$\square$

3.2.3 Discussion of Exact Results

s:int-dis

For the interval problem intuition suggests that learning with side information should be about the same as the standard learning with twice as large a sample. Indeed, both $x_1, \ldots, x_m$ and $s(x_1), \ldots, s(x_m)$ are i.i.d. samples from the uniform distribution. However, this intuition ignores the fact that the complete sample $x_1, \ldots, x_m, s(x_1), \ldots, s(x_m)$ is highly dependent. In the interval problem we see the effect of non-i.i.d. data clearly. It corresponds to the overlap phenomenon: the learning rate depends on how much the target T and its side information translate T' overlap each other. As a corollary of results obtained from exact calculations we state the following observation.

t:si-disc

Theorem 3.2.12 (Corollary of Theorems 3.2.1 and 3.2.7). *Let $T = (a, b)$ with $\ell = b - a$. Then for m large*

$$\frac{P^m\{\bar{x} \in X^m \; ; \; d_P((a,b), \widetilde{h}_m(\bar{x})) > \epsilon\}}{P^m\{\bar{x} \in X^m \; ; \; d_P((a,b), h_m(\bar{x})) > \epsilon\}}$$

$$= \begin{cases} O\left(\frac{1-2\epsilon}{1-\epsilon}\right)^{m-1}, & \text{if } \ell \in (\epsilon, \frac{1}{2}] \cup (\frac{1}{2}+\epsilon, 1], \\ O\left(\frac{2(\ell-\epsilon)}{1-\epsilon}\right)^{m-1}, & \text{if } \ell \in (\frac{1}{2}, \frac{1}{2}+\frac{\epsilon}{2}), \\ O\left(\frac{2(1-\ell)}{1-\epsilon}\right)^{m-1}, & \text{if } \ell \in [\frac{1}{2}+\frac{\epsilon}{2}, \frac{1}{2}+\epsilon]. \end{cases}$$

In particular, for T such that $\ell = \frac{1}{2} + \frac{\epsilon}{2}$ and m large

$$\frac{P^m\{\bar{x} \in X^m \; ; \; d_P((a,b), \widetilde{h}_m(\bar{x})) > \epsilon\}}{P^m\{\bar{x} \in X^m \; ; \; d_P((a,b), h_m(\bar{x})) > \epsilon\}} = O\left(\frac{1}{m}\right).$$

We see that typically side information yields to exponential improvement in the learning speed. However, the improvement for large samples fails in a special case when $T = (a, b)$ with $b - a = \frac{1}{2} + \frac{\epsilon}{2}$. The improvement rate will also depend on the underlying true concept. Thus, a worst case analysis will not in general capture a good picture of side information improvement, and some measure of average improvement rate is preferable. One can expect that overlap phenomena will become increasingly important as the amount of side information increases.

Next we discuss the results in more detail. Recall that the algorithm forms the estimate based on how close to the endpoints the positive samples fall. First if $T = (a, b)$ such that $b-a < 1/2$, if a sample falls in T, its side information translate falls outside T and similarly for the pair $s(x)$ and x. The probability of obtaining positive samples (original or side information) is $2(b-a)$. In this case we obtained $P^m\{\bar{x} \in X^m \ ; \ d_P(T, h_m) > \epsilon\} = (1-2\epsilon)^m + 2m\epsilon(1-2\epsilon)^{m-1}$. Even though this is not the learning rate with a $2m$ sample, $(1-\epsilon)^{2m} + 2m\epsilon(1-\epsilon)^{2m-1}$, those two differ very little.

Similar phenomenon occurs if T and T' overlap by more than ϵ. In this case, both a sample and its translate may be positive, but the algorithm is unlikely to form an estimate, say, $(x, s(x))$. This is because a sample close to one endpoint produces a translate that is not close to the other endpoint. As an extreme case, the learning algorithm is likely to operate on dependent data when $b-a = 1/2 + \epsilon/2$. Then a sample close to one endpoint yields a translate close to the other endpoint. In this case we saw that the large sample exponential improvement from side information disappears. On the other hand, if $b-a \in (1/2, 1/2+\epsilon)$ it is possible that $d_p(T, \widetilde{h}_m) < \epsilon$ even with just one sample. Thus $\mathrm{Err}(T, \widetilde{h}_m, m, \epsilon)$ decreases fast for small m, see Figure 3.4 for illustration of exact learning rates.

Remark 3.2.13. An upper bound for $P^m\{\bar{x} \in X^m \ ; \ d_P((0,b), \widetilde{h}_m) > \epsilon\}$ that holds for all b can be obtained easily as follows. Let $I_1 = (0, \frac{\epsilon}{2})$, $I_2 = (b - \frac{\epsilon}{2}, b)$, $I_3 = (\frac{1}{2}, \frac{1}{2} + \frac{\epsilon}{2})$ and $I_4 = (b' - \frac{\epsilon}{2}, b')$, where $b' = (b + 1/2) \mod 1$. Then the event {data in $I_1 \cup I_3$ and data in $I_2 \cup I_4$ } $\subset \{d_P(T, \widetilde{h}_m) \leq \epsilon\}$, and by taking complements

$$P^m\{\bar{x} \in X^m \ ; \ d_P(T, \widetilde{h}_m) > \epsilon\}$$
$$= Pr\{\neg\{\text{data in } I_1 \cup I_3 \text{ and data in } I_2 \cup I_4\}\} \leq 2(1-\epsilon)^m.$$

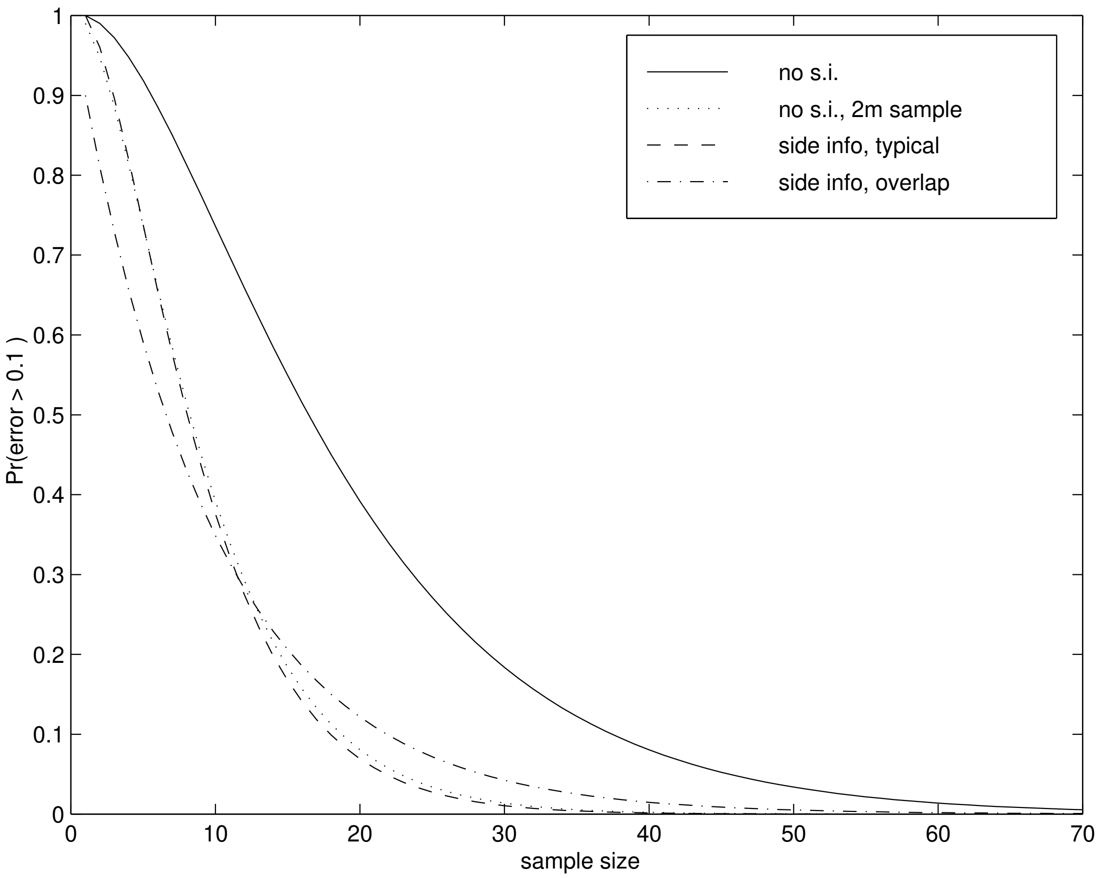

Figure 3.4: Illustration of exact learning rates from Theorems 3.2.1 and 3.2.7 with $\epsilon = 0.1$. Solid line is $\mathrm{Err}(T, h_m, m, \epsilon)$ and dotted line corresponds to learning with twice as large a sample without side information. Dashed line is the learning rate with side information for targets $T = (a, b)$ with $b - a \in (\epsilon, 1/2] \cup (1/2 + \epsilon, 1]$ and dash-dot line is the learning rate with side information when $b - a = 1/2 + \epsilon/2$.

The above estimate shows that

$$\lim_{m\to\infty} \frac{P^m\{\bar{x} \in X^m\ ;\ d_P(T, \widetilde{h}_m) > \epsilon\}}{P^m\{\bar{x} \in X^m\ ;\ d_P(T, h_m) > \epsilon\}} = 0.$$

Remark 3.2.14. We have calculated learning speed when the misclassification error of the estimate is greater than ϵ. In order to generalize from these results it is more convinient to consider estimates with misclassification error ϵ or larger. By taking $\int_a^b g(x)dx \geq \epsilon$ we obtain $P_g^m\{\bar{x} \in X^m\ ;\ d_{P_g}(T, h_m) \geq \epsilon\} = (1-\epsilon)^m + m\epsilon(1-\epsilon)^{m-1}$ and similarly with side information the results are the same with $P^m\{\bar{x} \in X^m\ ;\ d_P(T, \widetilde{h}_m) \geq \epsilon\}$ with the exception $\ell = b - a = 1/2 + \epsilon$ and in this case $P^m\{\bar{x} \in X^m\ ;\ d_P((a,b), \widetilde{h}_m) \geq \epsilon\} = (1-2\epsilon)^m + 2m\epsilon(1-2\epsilon)^{m-1}$.

3.3 Asymptotic Improvement Rates from Large Deviations

s:largedev

In this section we focus on the exponential improvement from side information we indicated in the interval problem. We assume that the data is drawn from non-uniform distribution P_g with density g and we consider more general side information mappings. However, our results are asymptotic log learning rates. Given a target $T = (a, b)$ we indicate the asymptotic exponential advantage of side information.

The large deviations theory studies asymptotic decay rates in probabilities of significant deviations from the mean in large number laws and ergodic limits. Further, the exponential decay rates are computable in terms of entropy functions. This will allow us to study the asymptotic advantage of the side information in a more general setting. By using the large deviations techniques we study the rate at which

$$\frac{1}{m} \ln P_g^m\{\bar{x} \in X^m\ ;\ d_{P_g}(T, \widetilde{h}_m) \geq \epsilon\}$$

converges to zero as $m \to \infty$.

This section is organized as follows: we begin by stating the large deviations theorem. The standard large deviations lower bound is not strong enough for our learning problem and in Section 3.3.1 we prove a better problem specific lower bound. Section 3.3.2 characterizes measures with the minimum relative entropy. By using machinery developed in these earlier sections we indicate the asymptotic exponential improvement for two learning algorithms in Sections 3.3.3 and 3.3.4.

We briefly recall the notation. Let $X_1, \ldots, X_m$ be an i.i.d. sample from distribution P_g with measurable density function g such that $g(x) > 0$ for Lebesgue almost every x in $[0,1]$. Let $T = (a,b)$ denote the target interval and let $T' = (\alpha, \beta) = (s^{-1}(a), s^{-1}(b))$ be the translate of the target under the side information mapping $s(x)$. The side information mapping $s(x)$ is assumed to be continuous, one-to-one, onto and $s(I) = I'$ where I and I' are intervals. All intervals are understood mod 1 and calculations are mod 1 too.

Given a learning algorithm h and an accuracy parameter ϵ, the idea is to characterize measures that lead to estimates with misclassification error ϵ or larger. This characterization is done in terms of the supports of the measures. Let $B = \{\mu \ ; \ h$ applied to μ has misclassification error $\geq \epsilon\}$. We can then show that

$$P_g^m\{\bar{x} \in X^m \ ; \ L^m \in B\} = P_g^m\{\bar{x} \in X^m \ ; \ d_{P_g}(T, h_m) \geq \epsilon\},$$

where $L^m = 1/m \sum_{i=1}^m \delta_{X_i}$ is the empirical measure. Our main conclusion is the improvement due to side information in the exponential decay rate. However, the improvement is related directly to P_g, the side information mapping and to the learning algorithm. Techniques of the proofs are geometric and of broader applicability in similar problems.

We quote the large deviations theorem for empirical measures from [20]. The topology in the space of probability measures $\mathcal{M}$ is the topology of weak convergence. Note that all large deviation upper and lower bounds to be stated do not depend on the fact that P_g has a density. We just write P_g in place of a general measure for the ease of notation and the future application in mind.

Theorem 3.3.1 (Large deviations theorem [20]). *Let* $X_1, X_2, \ldots$ *be a sequence of i.i.d. random variables taking values in* $\mathbb{R}$, $L^m = 1/m \sum_{i=1}^m \delta_{X_i}$ *and let* $P_g \in \mathcal{M}(\mathbb{R})$ *be the distribution of* X_1. *For* $\nu \in \mathcal{M}(\mathbb{R})$, *define the relative entropy of* ν *with respect to* P_g *by the formula*

e:relent

$$R(\nu \,\|\, P_g) = \begin{cases} \int \ln \frac{d\nu}{dP_g} d\nu, & \text{if } \nu << P_g \text{ and } \int |\ln \frac{d\nu}{dP_g} d\nu| < \infty, \\ \infty, & \text{otherwise.} \end{cases} \tag{3.3.4}$$

Then

e:ldub

$$\limsup_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in K\} \leq -\inf_{\nu\in K} R(\nu \,\|\, P_g), \quad \text{for each closed set } K \tag{3.3.5}$$

and

$$\liminf_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in G\} \geq -\inf_{\nu\in G} R(\nu \,\|\, P_g), \quad \text{for each open set } G.$$

$R(\nu \,\|\, P_g)$ is a convex function of ν and it measures the discrepancy between ν and P_g in the sense that $R(\nu \,\|\, P_g) \geq 0$ with equality if and only if $\nu = P_g$.

We will use the large deviations upper bound (3.3.5) with $\bar{B}$, the closure of B in the weak topology, in the place of K. Unfortunately, in examples we consider $\overset{\circ}{B}$, the interior of B, is empty. However, the definition of the set B in terms of supports of the measures allows us to develop a better large deviation lower bound as a stochastic control problem, yielding a bound

$$\liminf_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in B\} \geq -\inf_{\nu\in B} R(\nu \,\|\, P_g). \tag{3.3.6}$$

For the learning algorithms we discuss (simple interior approximation and general averaging algorithms) the bounds (3.3.5) and (3.3.6) actually match. This enables us to study the asymptotic advantage of the side information. In most cases it is easy to characterize this advantage. However, by comparing to exact calculations in a special case, we will see that large deviations method will only detect the exponential advantage of the side information.

3.3.1 Large Deviations Lower Bound

s:ldlb

We derive the large deviations lower bound via a stochastic control problem based on [19].

Assume that a set of measures B has the following property:

e:setb

$$\text{If } X_1, \ldots, X_m \text{ are drawn according to } \mu \in B, \text{ then } L^m = 1/m \sum_{i=1}^{m} \delta_{X_i} \in B. \tag{3.3.7}$$

l:ldlb

Lemma 3.3.2 (Large deviations lower bound). *Let $X_1, \ldots, X_m$ be i.i.d. with common distribution P_g, then*

e:ldlb

$$\liminf_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in B\} \geq -\inf_{\gamma \in B} R(\gamma \,\|\, P_g), \tag{3.3.8}$$

where B satisfies the property (3.3.7).

The rest of this section contains the proof of the above lemma. We begin by introducing the stochastic control setting and the stochastic representation theorem.

Define $\phi(\mu) = 0$, if $\mu \in B$ and $\phi(\mu) = \infty$ otherwise. Further, let $h_k(\mu) = k \wedge \phi(\mu)$ so that h_k is bounded and let

e:wkm

$$W_k^m := -\frac{1}{m} \ln E\Big[\exp\big(-m h_k(L^m)\big)\Big]. \tag{3.3.9}$$

The function W_k^m can be represented as a minimum cost function of an associated stochastic control problem. Let $\mathcal{M}_c = \{\mu \,|\, \mu(0,1) = c\}$, the set of measures with total mass equal to c and consider random subprobability measures (total mass may be less than 1) for $j = 1, \ldots, m$ such that $L_0^m = 0$ and $L_j^m = 1/m \sum_{k=1}^{j} \delta_{X_k}$. Each L_j^m takes values in $\mathcal{M}_{j/m}$. The state space at time i of the stochastic control problem is $\mathcal{M}_{i/m}$ and the control space is $\mathcal{M}$, the space of probability measures on $(0,1)$. Let ν_i denote the control applied at time i and the dependence of ν_i on $\mu \in \mathcal{M}_{i/m}$ is denoted by $\nu_i = \nu_i(\cdot \,|\, \mu)$. The controls are feedback controls so $\nu_i(\cdot \,|\, \mu)$ is a stochastic kernel given $\mathcal{M}_{i/m}$, i.e., for every Borel subset E the function mapping $\mu \mapsto \nu_i(E \mid \mu) \in [0,1]$ is measurable. The running cost for control ν is taken to be $1/m\, R(\nu \,\|\, P_g)$ and the terminal cost is h_k. Applying controls ν_j^m, $j = 1, \ldots, m$ gives rise to controlled random variables $\bar{X}_j^m$ and controlled empirical subprobability measures $\bar{L}_j^m$. $\bar{X}_j^m$ has the measure $\nu_j(\cdot \mid \bar{L}_j^m)$. The expectation, in the space where $\bar{X}_j^m$ and $\bar{L}_j^m$ are defined, is denoted by $\bar{E}$.

Define quantities $W_k^m(i,\mu)$ analogous to W_k^m such that for each $i \in \{0, \ldots, m\}$ and each measure $\mu \in \mathcal{M}_{i/m}$ we condition on the event $\{\bar{L}_i^m = \mu\}$. Let

$$\begin{aligned} W_k^m(i,\mu) &:= \frac{1}{m} \ln E_{i,\mu}\Big[\exp\big(-mh_k(\bar{L}^m)\big)\Big] \\ &= \frac{1}{m} \ln E\Big[\exp\big(-mh_k(\bar{L}^m)\big) \mid \bar{L}_i^m = \mu\Big]. \end{aligned}$$

It can be shown that $W_k^m(i,\mu)$ satisfies the dynamic programming equation

$$W_k^m(i,\mu) = \inf_\nu \{\frac{1}{m} R(\nu \,\|\, P_g) + \int W_k^m(i+1, \mu + \frac{1}{m}\delta_y)\nu(dy)\}$$

with the terminal condition

$$W_k^m(m,\mu) = h_k(\mu).$$

Further, the W_k^m defined by (3.3.9) is the unique solution to the dynamic programming problem. We arrive to the stochastic representation theorem:

Theorem 3.3.3 ([19, p. 56]). *Let h_k be a bounded measurable function mapping $\mathcal{M}$ into $\mathbb{R}$. Then for all $m \in \mathbb{N}$,*

$$W_k^m = \inf_{\{\nu_j^m\}} \bar{E}\left[\frac{1}{m}\sum_{j=1}^m R\big(\nu_j^m(\cdot|\bar{L}_j^m) \,\|\, P_g\big) + h_k(\bar{L}^m)\right]. \tag{3.3.10}$$

Proposition 3.3.4. *Assume a set B has the property* (3.3.7)*. Then*

$$\limsup_{m\to\infty} W_k^m \le \inf_{\mu \in B} R(\mu \,\|\, P_g).$$

Proof. Let $\gamma \in B$ and for $j = 1, \ldots, m$ define $\nu_j^m = \gamma$. The representation theorem (3.3.10) gives that

$$W_k^m \le \bar{E}\big[R(\gamma \,\|\, P_g) + h_k(\bar{L}^m)\big].$$

With this choice of control $\bar{L}^m$ is the empirical measure of i.i.d. random variables with distribution γ. As $\gamma \in B$, also $\bar{L}^m \in B$, and $h_k(\bar{L}^m) = 0$ for all m. Hence $\limsup_{m\to\infty} W_k^m \le R(\gamma \,\|\, P_g)$. As $\gamma \in B$ was arbitrary,

$$\limsup_{m\to\infty} W_k^m \le \inf_{\gamma\in B} R(\gamma \,\|\, P_g).$$

□

Proof of the large deviations lower bound, Lemma 3.3.2.
Observe that $E\big[\exp(-mh_k(L^m))\big] = e^{-mk} Pr\{L^m \in B^c\} + Pr\{L^m \in B\} \le e^{-mk} + Pr\{L^m \in B\}$. Thus

$$\ln E\Big[\exp\big(-mh_k(L^m)\big)\Big] \le \begin{cases} \ln(2) + \ln(e^{-mk}), & \text{if } e^{-mk} \ge Pr\{L^m \in B\}, \\ \ln(2) + \ln Pr\{L^m \in B\}, & \text{otherwise.} \end{cases}$$

Dividing by m and taking liminf gives us

$$\begin{aligned} \max&\left(\liminf_{m\to\infty} \frac{1}{m} Pr\{L^m \in B\}, -k\right) \\ &\ge \liminf_{m\to\infty} \frac{1}{m} \ln E\Big[\exp\big(-mh_k(L^m)\big)\Big] \\ &= \liminf_{m\to\infty} -W_k^m = -\limsup_{m\to\infty} W_k^m \ge -\inf_{\gamma\in B} R(\gamma \,\|\, P_g), \end{aligned}$$

where the last inequality follows from Proposition 3.3.4. Let $k \to \infty$ to obtain

$$\liminf_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in B\} \ge -\inf_{\gamma\in B} R(\gamma \,\|\, P_g).$$

□

3.3.2 The Measure with Minimum Relative Entropy

s:minent

The large deviation method determines the exponential convergence rates in terms of the relative entropy. If the large deviation upper and lower bounds match, the asymptotic rate is given by calculating $\inf_{\mu \in B} R(\mu \,\|\, P_g)$. Recall that the idea was to define the set of "bad" measures B in terms of the supports of the measures.

Among those measures that have support $S \subset (0,1)$, the one with minimum relative entropy with respect to P_g has a density which is a scaled version of the density $g(x)$.

p:minent

Proposition 3.3.5.

$$\inf_{\operatorname{supp}\mu = S} R(\mu \,\|\, P_g) = R(P_f \,\|\, P_g) = -\ln \int_S g(x)dx,$$

where P_f has density $f(x) = \frac{g(x)}{\int_S g(x)dx} \mathbf{1}_S(x)$.

Proof. Let μ have support S. Recall that $R(\mu \,\|\, P_g) = \infty$ if μ is not absolutely continuous with respect to P_g. Let $dP_f/dP_g = (\int_S g(x)dx)^{-1} \mathbf{1}_S(x)$. Then P_g is absolutely continuous with respect to P_f. Further, if μ is absolutely continuous with respect to P_g, then also μ is absolutely continuous with respect to P_f. Now the result follows from the chain rule of the Radon-Nikodym derivatives.

$$\begin{aligned} R(\mu \,\|\, P_g) = \int_S \ln \frac{d\mu}{dP_g} d\mu = \int_S \ln \frac{d\mu}{dP_f}\frac{dP_f}{dP_g} d\mu = \int_S \ln \frac{d\mu}{dP_f} d\mu + \int_S \ln \frac{dP_f}{dP_g} d\mu \\ = R(\mu \,\|\, P_f) - \ln \int_S dP_g \geq -\ln \int_S dP_g, \end{aligned}$$

where the last equality is obtained by substituting dP_f/dP_g. The last inequality is actually an equality if $\mu = P_f$, because $R(\mu \,\|\, P_f) = 0$ if and only if $\mu = P_f$. □

3.3.3 Approximating the Target from Inside

s:ld-int-alg

We first illustrate the large deviations method for an algorithm that approximates the unknown target interval from inside. The estimate for the target is the smallest interval containing positive (original or side information) samples. In this simple case we can compare results to the exact calculations when P_g is the uniform distribution and the side information mapping is the translation by $1/2$. We begin this section by stating the main result. Its proof is given at the end of the section, after necessary definitions and propositions.

For learning without side information the exact learning rate calculations show that

$$\lim_{m\to\infty}\frac{1}{m}\ln P_g^m\{\bar{x}\in X^m\ ;\ d_{P_g}(T,h_m)\geq\epsilon\}=\ln(1-\epsilon).$$

In general, the advantage of side information comes from the fact that the random samples have to avoid not only the endpoints of the target interval but also the endpoints α and β of the translated interval. The advantage of side information is indicated by the following theorem.

t:ld-impr

Theorem 3.3.6. *Let $\widetilde{h}_m$ be the interior approximation of the target with side information. Assume that samples are from distribution P_g with measurable density function g such that $g(x)>0$ for Lebesgue almost every x in $[0,1]$ and the side information mapping $s(x)$ is continuous, one-to-one, onto and $s(I)=I'$ where I and I' are intervals. Then*

$$\lim_{m\to\infty}\frac{1}{m}\ln P_g^m\{\bar{x}\in X^m\ ;\ d_{P_g}((a,b),\widetilde{h}_m)\geq\epsilon\}=\ln(1-\epsilon-\epsilon_1),$$

where

$$\epsilon_1=\inf\left\{\int g(x)\mathbf{1}_{(s^{-1}(a,c)\cup s^{-1}(d,b))\setminus((a,c)\cup(d,b))}dx\ ;\right.$$
$$\left. a<c\leq d<b,\int g(x)\mathbf{1}_{(a,c)\cup(d,b)}\,dx=\epsilon\right\}.$$

When $\epsilon_1>0$ large deviations shows a strict improvement over the learning rate without side information. For example, if we consider the case when the target interval T and its side information translate T' do not overlap and if P_g is the uniform distribution, we get by calculating the minimum relative entropy that

$$\lim_{m\to\infty}\frac{1}{m}\ln P_g^m\{\bar{x}\in X^m\ ;\ d_{P_g}(T,\widetilde{h}_m)\geq\epsilon\}=\ln(1-2\epsilon),$$

which is in agreement with the exact calculations carried out before.

The term $(s^{-1}(a,c) \cup s^{-1}(d,b)) \setminus ((a,c) \cup (d,b))$ takes into account that in some cases avoiding the endpoints a and b of the target interval automatically means avoiding the endpoints of the translated interval as well and $\epsilon_1 = 0$. This is exactly what happens when the length of the target interval is $\epsilon/2 + 1/2$ in the example of learning under the uniform distribution and side information mapping $s(x) = (x + 1/2)$ mod 1. However, Theorem 3.2.12 shows the improvement of $1/m$ for m large, and the large deviations method fails to capture this. This does not contradict the exact calculations, because in this case $\lim_{m\to\infty} 1/m \ln P^m\{\bar{x} \in X^m \; ; \; d_P(T, \widetilde{h}_m) > \epsilon\} = \lim_{m\to\infty} 1/m \ln((1-\epsilon)^m) = \ln(1-\epsilon)$. The improvement from the side information has to be exponential in order to be detected by the large deviations technique.

To build machinery to prove Theorem 3.3.6 we begin by defining the set of measures yielding to an estimate with misclassification error ϵ or larger. If the target interval is $T = (a,b)$ then its side information translate is $T' = (\alpha, \beta) = (s^{-1}(a), s^{-1}(b))$ and the estimate is formed based on how close the positive samples fall to the endpoints of either T or T'. The closeness is going to be defined in terms of the following support parameters. Given a probability measure μ, define support parameters

$$
\begin{aligned}
a_r &= \inf\{a < t \le b,\ \mu\,(a,t] > 0\},\\
b_\ell &= \sup\{a \le t < b,\ \mu\,[t,b) > 0\},\\
\alpha_r &= \inf\{\alpha < t \le \beta,\ \mu\,(\alpha,t] > 0\},\\
\beta_\ell &= \sup\{\alpha \le t < \beta,\ \mu\,[t,\beta) > 0\},
\end{aligned}
$$

where subscripts r and ℓ refer to the right and to the left, respectively. Now $(a, a_r) \cup (b_\ell, b) \cup (\alpha, \alpha_r) \cup (\beta_\ell, \beta) \subset (\operatorname{supp}\mu)^c$, where $\operatorname{supp}\mu$ stands for the support of μ.

Let $\epsilon > 0$. Support parameters $a_r, b_\ell, \alpha_r, \beta_\ell$ are called ϵ-admissible if

$$
\int_{T \setminus \widehat{T}(a_r, b_\ell, \alpha_r, \beta_\ell)} g\, dx \ge \epsilon,
$$

where $T \setminus \widehat{T}(a_r, b_\ell, \alpha_r, \beta_\ell)$ equals

$$\big(a, a + (a_r - a) \wedge (s(\alpha_r) - a)\big) \cup \big(b - (b - b_\ell) \wedge (b - s(\beta_\ell)), b\big).$$

Define

e:bad-measure
$$B_\epsilon := \{\mu \,;\ \text{ support parameters of } \mu \text{ are } \epsilon\text{-admissible } \} \tag{3.3.11}$$

If the learning algorithm is the approximation from inside, B_ϵ is the set that leads to an estimate with misclassification error ϵ or larger, i.e., if $L_m \in B_\epsilon$ the estimate $\widetilde{h}_m$ has error ϵ or larger. Note that the set of ϵ-admissible support parameters is closed in $\mathbb{R}^4$.

In order to calculate the large deviations upper bound for log learning rates we need the following proposition:

Proposition 3.3.7. *The set B_ϵ is closed in the topology of weak convergence.*

Proof. Let $\{\mu_n\} \in B_\epsilon$ with support parameters $a_r^n, b_\ell^n, \alpha_r^n$ and β_ℓ^n and let μ_n converge weakly to μ with support parameters $a_r, \beta_\ell, \alpha_r$ and β_ℓ. We show that $\mu \in B_\epsilon$.

Now for all $\epsilon' > 0$ there exists N such that if $n > N$

$$\begin{aligned} a_r^n &\le a_r + \epsilon' & \qquad \alpha_r^n &\le \alpha_r + \epsilon' \\ b_\ell^n &\ge b_\ell - \epsilon' & \qquad \beta_\ell^n &\ge \beta_\ell - \epsilon'. \end{aligned}$$

This can be seen by assuming the converse, i.e., there exist an $\epsilon' > 0$ such that for each N there is $n > N$ with $a_r^n > a_r + \epsilon'$. Thus there exists a subsequence n_j with the above property. For the ease of notation, the subscript is dropped. Take a function h as in Figure 3.5. Then $\int h d\mu > 0$ but $\int h d\mu_n = 0$ so $\lim_{n\to\infty} \int h d\mu_n \neq \int h d\mu$ contradicting the fact that $\{\mu_n\}$ converges weakly to μ. Because $a_r^n, b_\ell^n, \alpha_r^n$ and β_ℓ^n are admissible

$$\epsilon \le \int \mathbf{1}_{T\setminus\widehat{T}(a_r^n, b_\ell^n, \alpha_r^n, \beta_\ell^n)} g\, dx \le \int \mathbf{1}_{T\setminus\widehat{T}(a_r+\epsilon', b_\ell-\epsilon', \alpha_r+\epsilon', \beta_\ell-\epsilon')} g\, dx,$$

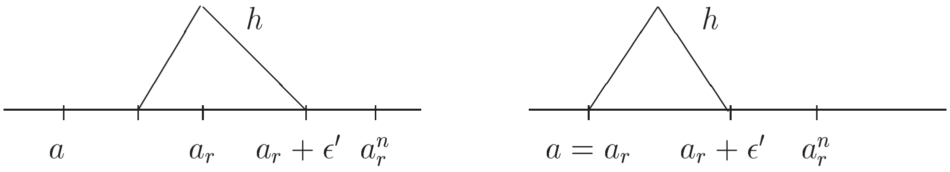

Figure 3.5: The function h used in the proof.

f:h

for all $\epsilon' > 0$. But letting ϵ' approach to zero, we get

$$\int \mathbf{1}_{T \setminus \widehat{T}(a_r, b_\ell, \alpha_r, \beta_\ell)} g \, dx \geq \epsilon,$$

thus $\mu \in B_\epsilon$. $\square$

Proof of Theorem 3.3.6. The definition of B_ϵ implies that

$$Pr\{L^m \in B_\epsilon\} = P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}(T, \widetilde{h}_m) \geq \epsilon\}.$$

Now by the previous proposition the large deviation upper bound (3.3.5) and the lower bound (3.3.8) match, giving us

$$\begin{aligned} \lim_{m \to \infty} \frac{1}{m} \ln Pr\{L^m \in B_\epsilon\} &= \lim_{m \to \infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}(T, \widetilde{h}_m) \geq \epsilon\} \\ &= - \inf_{\mu \in B_\epsilon} R(\mu \,\|\, P_g). \end{aligned}$$

Let a_r, b_ℓ, α_r, and β_ℓ be support parameters that are ϵ-admissible. Proposition 3.3.5 gives that the measure with minimum relative entropy with respect to P_g with these fixed support parameters has density $g(x)(\int \mathbf{1}_{((a,a_r)\cup\cdots\cup(\beta_\ell,\beta))^c} g(x)dx)^{-1}$ and the minimum relative entropy is given by

$$\inf \left\{ -\ln \left(1 - \int_{(a,a_r)\cup\cdots\cup(\beta_\ell,\beta)} g(x)dx \right); a_r, \ldots, \beta_\ell \text{ are } \epsilon\text{-admissible} \right\}$$
$$= -\ln(1-c),$$

where

$$c = \inf \left\{ \int_{(a,a_r)\cup\cdots\cup(\beta_\ell,\beta)} g(x)dx; a_r, \ldots, \beta_\ell \text{ are } \epsilon\text{-admissible} \right\}.$$

By continuity we need to consider ϵ-admissible support parameters a_r, b_ℓ, α_r, and β_ℓ such that

$$\int \mathbf{1}_{T\setminus\widehat{T}(a_r,b_\ell,\alpha_r,\beta_\ell)} g(x)\,dx = \epsilon.$$

If $\int \mathbf{1}_{T\setminus\widehat{T}(a_r,b_\ell,\alpha_r,\beta_\ell)} g(x)\,dx > \epsilon$, then, for example, $a_r - \rho, b_\ell, \alpha_r, \beta_\ell$ are ϵ-admissible for small enough ρ, and the corresponding relative entropy is smaller. Further, if a_r and b_ℓ are such that

$$\int_{(a,a_r)\cup(b_\ell,b)} g(x)\,dx > \epsilon$$

we can again reduce the relative entropy by taking a_r and b_ℓ closer to a and b, respectively.

Thus we assume that

$$\int_{(a,a_r)\cup(b_\ell,b)} g(x)\,dx = \epsilon \qquad \text{and} \qquad \int \mathbf{1}_{T\setminus\widehat{T}(a_r,b_\ell,\alpha_r,\beta_\ell)} g(x)\,dx = \epsilon.$$

Note that now ϵ-admissibility implies that

$$s^{-1}(a,a_r) \cup s^{-1}(b_\ell,b) \subseteq (\alpha,\alpha_r) \cup (\beta_\ell,\beta)$$

and the relative entropy is minimized when

$$s^{-1}(a,a_r) \cup s^{-1}(b_\ell,b) = (\alpha,\alpha_r) \cup (\beta_\ell,\beta).$$

Thus

$$\begin{aligned}
&\inf\Big\{ \int_{(a,a_r)\cup\cdots\cup(\beta_\ell,\beta)} g(x)dx;\, a_r,\ldots,\beta_\ell \text{ are } \epsilon\text{-admissible}\Big\} \\
&\quad= \inf\Big\{ \int_{(a,a_r)\cup\cdots\cup(b_\ell,b)} g(x)dx + \int_{\substack{((\alpha,\alpha_r)\cup\cdots\cup(\beta_\ell,\beta)) \\ \setminus((a,a_r)\cup\cdots\cup(b_\ell,b))}} g(x)dx;\, a_r,\ldots,\beta_\ell \text{ are } \epsilon\text{-admissible}\Big\} \\
&\quad= \inf\Big\{ \epsilon + \int_{\substack{(s^{-1}(a,c)\cup s^{-1}(d,b)) \\ \setminus((a,c)\cup(d,b))}} g(x)\,dx;\, a<c\le d<b, \int_{(a,c)\cup(d,b)} g(x)\,dx = \epsilon\Big\} \\
&\quad= \epsilon + \epsilon_1,
\end{aligned}$$

where

$$\epsilon_1 = \inf\left\{ \int \mathbf{1}_{\substack{(s^{-1}(a,c)\cup s^{-1}(d,b))\\ \setminus((a,c)\cup(d,b))}} g(x)\,dx;\, a < c \le d < b, \int_{(a,c)\cup(d,b)} g(x)\,dx = \epsilon \right\}$$

and

$$\inf_{\mu\in B_\epsilon} R(\mu \,\|\, P_g) = -\ln(1-\epsilon-\epsilon_1).$$

$\square$

3.3.4 General Averaging Algorithms

s:ld-gen-alg

In this section we generalize the learning algorithm from interior approximation to an algorithm that forms its estimate based on how close samples (now positive and negative) fall to the end points of the target T or the translate $T' = (\alpha, \beta)$.

Given a probability measure μ on $(0, 1)$ define support parameters

$$a_\ell = \sup\{t \le a\,;\ \mu[t, a) > 0\},$$

$$a_r = \inf\{t \ge a\,;\ \mu(a, t] > 0\},$$

$$b_\ell = \sup\{t \le b\,;\ \mu[t, b) > 0\},$$

$$b_\ell = \inf\{t \ge b\,;\ \mu(b, t] > 0\},$$

and α_ℓ, α_r, β_ℓ, and β_r similarly.

Assume that we have two averaging algorithms f_a and f_b such that $f_a(x_1, x_2) \in [x_1, x_2]$, $f_b(x_1, x_2) \in [x_1, x_2]$ and f_a, f_b are continuous on x_1 and x_2. The estimate for the target interval $T = (a, b)$ is (f_a, f_b), where

$$f_a = f_a(a_\ell \vee s(\alpha_\ell), a_r \wedge s(\alpha_r)),$$

$$f_b = f_b(b_\ell \vee s(\beta_\ell), b_r \wedge s(\beta_r)),$$

$x \wedge y = \min\{x, y\}$ and $x \vee y = \max\{x, y\}$. Note that the estimate (f_a, f_b) depends continuously on a_ℓ, a_r, b_ℓ, b_r, α_ℓ, α_r, β_ℓ, and β_r. The idea in forming the estimate is to move the support parameters $\alpha_\ell, \ldots, \beta_r$ back to the target and consider support parameters that are closest (on the left and on the right) to a and b.

The advantage of the side information is not as easy to characterize for the general averaging algorithms as it was for the algorithm of a simple approximation from inside, where the estimate came directly from the values of the support parameters. Here for generalized averaging algorithms the learning without side information gives

$$\lim_{m\to\infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}(T, h_m) \geq \epsilon\} = \ln(1-\epsilon_2) \leq \ln(1-\epsilon),$$

where

$$\epsilon_2 = \inf\left\{\int_{(a_\ell,a_r)\cup(b_\ell,b_r)} g(x)\,dx \; ; \; a_\ell, a_r, b_\ell, b_r \text{ satisfy } (3.3.12)\right\},$$

e:ave-ent

$$\int \mathbf{1}_{(f_a(a_\ell,a_r),f_b(b_\ell,b_r))\Delta(a,b)} g(x)\,dx = \epsilon. \tag{3.3.12}$$

The reason for the inequality is that the estimate for the target h_m may not coincide with the values of the support parameters, like in the simple approximation algorithm. For general averaging algorithms without side information, for example, $f_a(a_\ell, a_r) \in [a_\ell, a_r]$.

We can show that the procedure of Theorem 3.3.6 can be repeated for the averaging algorithm and we get:

t:ld-impr1

Theorem 3.3.8. *Let $\widetilde{h}_m$ be a general averaging algorithm with side information. Assume that samples are from distribution P_g with measurable density function g such that $g(x) > 0$ for Lebesgue almost every x in $[0,1]$ and the side information mapping $s(x)$ is continuous, one-to-one, onto and $s(I) = I'$ where I and I' are intervals. Then*

$$\lim_{m\to\infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}((a,b), \widetilde{h}_m) \geq \epsilon\} = \ln(1-\epsilon_3),$$

where

$$\epsilon_3 = \inf \Big\{ \int_{(a_\ell,a_r)\cup(b_\ell,b_r)} g(x)\,dx \\ + \int_{\substack{(s^{-1}(a_\ell,a_r)\cup s^{-1}(b_\ell,b_r)) \\ \setminus((a_\ell,a_r)\cup(b_\ell,b_r))}} g(x)\,dx; a_\ell, a_r, b_\ell, b_r \text{ satisfy } (3.3.12) \Big\}.$$

Again, if $((a_\ell, a_r) \cup (b_\ell, b_r)) = (s^{-1}(a_\ell, a_r) \cup s^{-1}(b_\ell, b_r))$ the advantage of side information can not be detected.

Now we indicate that although the learning algorithm is different from the interior approximation, similar type of process can be carried out to obtain the log learning rates with side information. In this case our interest is focused on the large deviation upper bounds.

We begin again by defining the set of measures that yield a poor estimate with a generalized averaging algorithm. For $\epsilon > 0$, define

e:Bavealg

$$B_\epsilon = \Big\{ \mu \mid \mu \text{ has support parameters } a_\ell, \dots, \beta_r \\ \text{such that } \int_{(f_a,f_b)\Delta(a,b)} g(x)\,dx \geq \epsilon \Big\}, \tag{3.3.13}$$

where $(f_a, f_b)\Delta(a, b) = ((f_a, f_b) \cap (a, b)^c) \cup ((f_a, f_b)^c \cap (a, b))$ is the symmetric difference of (f_a, f_b) and (a, b).

We have seen that if $f_a(x_1, x_2) = x_2$ and $f_b(x_1, x_2) = x_1$, the set B_ϵ is closed in the topology of weak convergence. For general averaging algorithms this is not necessarily the case: for example, if $f_a(x_1, x_2) = f_b(x_1, x_2) = 1/2\,(x_1 + x_2)$ one can find a sequence in B_ϵ that converges weakly to a measure for which the estimation error $\int_{(f_a,f_b)\Delta(a,b)} g(x)\,dx = 0$ and hence the measure is not in B_ϵ. It is possible to have point masses so that the support parameters are not symmetrically around the endpoints a, b, α, and β. Weak converge allows the point masses to die out, and the resulting measure has support parameters symmetrically around the endpoints. More formally, assume that a, b, α and β are suitable (so that D intervals do not overlap) and D is such that

$$\int_a^{a+D/3} g(x)\,dx + \int_{b-D/3}^{b} g(x)\,dx \geq \epsilon.$$

Consider a measure μ_n with $(\operatorname{supp}\mu_n)^c = (I_a \setminus \{a-\delta\}) \cup (I_b \setminus \{b+\delta\}) \cup (I_\alpha \setminus \{\alpha-\delta\}) \cup (I_\beta \setminus \{\beta+\delta\})$, where $I_x = (x-D, x+D)$ and let μ_n have $1/n$ masses at points $a-\delta$, $b+\delta$, $\alpha-\delta$, and $\beta+\delta$. We take δ small compared to D so that $f_a(a-\delta, a+D) > a + D/3$, similarly for f_b, and

$$\int_{(f_a, f_b)\Delta(a,b)} g(x)\,dx \geq \epsilon.$$

However, μ_n converges weakly to a measure μ with support parameters $a_\ell = a - D$, $a_r = a + D$, $b_\ell = b - D, \ldots, \beta_r = \beta + D$. But for those $(f_a, f_b) = (a, b)$ and the estimation error is zero. Thus $\mu \notin B_\epsilon$.

The large deviation upper bound states that

$$\limsup_{m\to\infty} \frac{1}{m} \ln Pr\{L^m \in B_\epsilon\} \leq -\inf_{\mu \in \bar{B}_\epsilon} R(\mu \,\|\, P_g),$$

where $\bar{B}_\epsilon$ is the closure of B_ϵ in the weak topology. However, even in this situation we will be able to get the upper and the lower bound to match. Namely, we show next that taking the closure does not yield a smaller infimum.

p:sameinf

Proposition 3.3.9. *For B_ϵ defined by* (3.3.13)

$$\inf_{\mu\in\bar{B}_\epsilon} R(\mu \,\|\, P_g) = \inf_{\mu \in B_\epsilon} R(\mu \,\|\, P_g).$$

Proof. We assume that $\inf_{\mu\in\bar{B}_\epsilon} R(\mu \,\|\, P_g) < \inf_{\mu\in B_\epsilon} R(\mu \,\|\, P_g)$ and derive a contradiction. Let δ be such that $\inf_{\mu\in B_\epsilon} R(\mu \,\|\, P_g) - \inf_{\mu\in\bar{B}_\epsilon} R(\mu \,\|\, P_g) > 2\delta$. Take $\mu_\delta \in \bar{B}_\epsilon$ so that $R(\mu_\delta \,\|\, P_g) < \inf_{\mu\in\bar{B}_\epsilon} R(\mu \,\|\, P_g) + \delta$. Necessarily $\mu_\delta \in \bar{B}_\epsilon \setminus B_\epsilon$, and let μ_δ have support parameters $a_\ell, a_r, \ldots, \beta_\ell, \beta_r$. Further, let μ_δ^g have the same support parameters and let

$$\frac{d\mu_\delta^g}{dP_g} = \left(\int \mathbf{1}_{((a_\ell,a_r)\cup\cdots\cup(\beta_\ell,\beta_r))^c}\, g(x)dx \right)^{-1}.$$

By Proposition 3.3.5 $R(\mu_\delta^g \,\|\, P_g) \le R(\mu_\delta \,\|\, P_g)$ and $R(\mu_\delta \,\|\, P_g) < R(\tilde\mu \,\|\, P_g)$ for all $\tilde\mu \in B_\epsilon$. Further by considering those measures in B_ϵ with density $cg(x)$, where c is a scaling constant, we have

$$\begin{aligned} R(\mu_\delta^g \,\|\, P_g) &= -\ln(1 - \int_{(a_\ell,a_r)\cup\cdots\cup(\beta_\ell,\beta_r)} g(x)dx) \\ &< -\ln(1 - \int_{(\tilde a_\ell,\tilde a_r)\cup\cdots\cup(\tilde\beta_\ell,\tilde\beta_r)} g(x)dx) = R(\tilde\mu \,\|\, P_g), \end{aligned}$$

where $\tilde\mu \in B_\epsilon$ has support parameters $\tilde a_\ell, \ldots, \tilde\beta_r$. Thus

e:intg(x)

$$\int_{(a_\ell,a_r)\cup\cdots\cup(\beta_\ell,\beta_r)} g(x)dx < \int_{(\tilde a_\ell,\tilde a_r)\cup\cdots\cup(\tilde\beta_\ell,\tilde\beta_r)} g(x)dx. \tag{3.3.14}$$

As $\mu_\delta \in \bar B_\epsilon$ there exists a sequence $\{\mu_n\} \in B_\epsilon$ converging weakly to μ_δ. Let μ_n have support parameters $a_\ell^n, a_r^n, \ldots, \beta_\ell^n, \beta_r^n$. We have shown previously that given $\epsilon' > 0$ there exists N so that whenever $n \ge N$, $a_\ell^n \in (a_\ell - \epsilon', a)$, $a_r^n \in (a, a_r + \epsilon'), \ldots, \beta_r^n \in (\beta, \beta_r + \epsilon')$. Now we show that actually $a_\ell^n \to a, \ldots, \beta_r^n \to \beta_r$.

Consider a sequence $\{\mu_n^g\} \in B_\epsilon$ with support $((a_\ell^n, a_r^n) \cup \cdots \cup (\beta_\ell^n, \beta_r^n))^c$ and the density $c_n g(x)$, where c_n is a scaling factor depending on the support parameters. Assume that $a_\ell^n \not\to a_\ell$, i.e., there exists a subsequence $a_\ell^{n_k}$ so that $a_\ell^{n_k} > a_\ell + \epsilon$ for all k. But then for k large

$$\int_{(a_\ell^{n_k}, a_r^{n_k})\cup\cdots\cup(\beta_\ell^{n_k},\beta_r^{n_k})} g(x)\,dx < \int_{(a_\ell,a_r)\cup\cdots\cup(\beta_\ell,\beta_r)} g(x)\,dx$$

contradicting (3.3.14) because $\mu_{n_k}^g \in B_\epsilon$. Hence $a_\ell^n \to a_\ell, \ldots, \beta_r^n \to \beta_r$. But then also $\{\mu_n^g\}$ converges weakly to μ_δ^g.

Let (f_a^n, f_b^n) be the estimate for the target T based on $a_\ell^n, \ldots, \beta_r^n$, and recall that f_a^n and f_b^n are continuous on $a_\ell^n, \ldots, \beta_r^n$. Further, the estimate (f_a, f_b) based on $a_\ell, \ldots, \beta_r$ satisfies

$$\int_{(f_a,f_b)\Delta(a,b)} g(x)\,dx < \epsilon.$$

But then for n large enough also

$$\int_{(f_a^n, f_b^n)\Delta(a,b)} g(x)\, dx < \epsilon,$$

which contradicts the fact that $\{\mu_n^g\} \in B_\epsilon$. □

Proof of Theorem 3.3.8. Proposition 3.3.9 implies that the upper and the lower bound for the large deviations agree, stating that

$$\lim_{m\to\infty} \frac{1}{m} \ln P_g^m\{\bar{x} \in X^m \; ; \; d_{P_g}((a,b), \widetilde{h}_m) \geq \epsilon\} = -\inf_{\mu\in B_\epsilon} R(\mu \,\|\, P_g).$$

Recall that a_ℓ, ... , β_r are ϵ-admissible if

$$\int_{(f_a, f_b)\Delta(a,b)} g(x)\, dx \geq \epsilon,$$

where

$$f_a = f_a(a_\ell \vee s(\alpha_\ell), a_r \wedge s(\alpha_r)),$$
$$f_b = f_b(b_\ell \vee s(\beta_\ell), b_r \wedge s(\beta_r)).$$

To calculate the minimum entropy we need to find

$$\inf \left\{ \int_{(a_\ell,a_r)\cup(b_\ell,b_r)\cup(\alpha_\ell,\alpha_r)\cup(\beta_\ell,\beta_r)} g(x)\, dx \; ; \; a_\ell, \ldots, \beta_r \text{ are } \epsilon\text{-admissible} \right\}.$$

By the continuity of f_a and f_b we need to consider support parameters $a_\ell, \ldots, \beta_r$ such that

$$\int_{(f_a, f_b)\Delta(a,b)} g(x)\, dx = \epsilon.$$

Let x_1, x_2, x_3, and x_4 be such that

$$P_g\Big(\big(f_a(x_1, x_2), f_b(x_1, x_2)\big)\Delta(a,b)\Big) = \epsilon,$$

where $x_1 = a_\ell \vee s(\alpha_\ell)$ and x_2, x_3, and x_4 similarly.

Corresponding support parameters have minimum relative entropy when $a_\ell = s(\alpha_\ell)$ and similarly for other support parameters. Thus

$$(\alpha_\ell, \alpha_r) \cup (\beta_\ell, \beta_r) = s^{-1}(a_\ell, a_r) \cup s^{-1}(b_\ell, b_r).$$

Also note that now the estimate with side information coincides with the one without side information. Thus we get

$$\begin{aligned}
&\inf\Big\{\int_{(a_\ell,a_r)\cup\cdots\cup(\beta_\ell,\beta_r)} g(x)\,dx;\, a_\ell,\ldots,\beta_r \text{ are } \epsilon\text{-admissible}\Big\}\\
&\quad= \inf\Big\{\int_{(a_\ell,a_r)\cup(b_\ell,b_r)} g(x)\,dx + \int_{\substack{((\alpha_\ell,\alpha_r)\cup(\beta_\ell,\beta_r))\\ \setminus((a_\ell,a_r)\cup(b_\ell,b_r))}} g(x)\,dx;\, a_\ell,\ldots,\beta_r \text{ are } \epsilon\text{-admissible}\Big\}\\
&\quad= \inf\Big\{\int_{(a_\ell,a_r)\cup(b_\ell,b_r)} g(x)\,dx\\
&\qquad\qquad + \int_{\substack{(s^{-1}(a_\ell,a_r)\cup s^{-1}(b_\ell,b_r))\\ \setminus((a_\ell,a_r)\cup(b_\ell,b_r))}} g(x)\,dx;\, a_\ell, a_r, b_\ell, b_r \text{ satisfy } (3.3.12)\Big\}.\\
&\quad= \epsilon_3.
\end{aligned}$$

Hence the minimum relative entropy is $-\ln(1-\epsilon_3)$. □

3.4 Numerical Results of Learning Open Intervals with Side Information

s:numres

We calculated analytically exact learning rates for the problem of learning an open interval under the uniform ditribution on $(0,1)$ with side information mapping $s(x) = (x+1/2) \bmod 1$. Whereas the learning rate without side information (the standard PAC problem) was independent of the density of the distribution, the proof of exact learning rates with side information can not be generalized to arbitrary distributions with densities. Moreover, the advantage of side information is highly distribution dependent. To illustrate the effect of side information on small samples for other distributions we rely on numerical simulations. In all the simulations the side information mapping is $s(x) = (x+1/2) \bmod 1$ and in first two examples the learning algorithm is the one used for analytical calculations, i.e., the smallest interval that is consistent with the observed data. The last simulated example is again with the uniform distribution on $(0, 1)$, but now the algorithm is different. All learning error calculations are done with $\epsilon = 0.1$.

We simulate learning rates as a function of the sample size. Given a new sample, an estimate is formed and its misclassification probability is calculated and the process continues by adding the next sample. More formally, let h_m and $\widetilde{h}_m$ be estimates for the target T without and with side information, respectively, and assume that $x_1, x_2, \ldots, x_M$ are i.i.d. random samples with the probability law P.

- Initialize the algorithms, i.e., form h_0 and $\widetilde{h}_0$.
- Repeat for $m = 1, \ldots, M$
    1. Observe a random sample x_m and update the estimate h_{m-1} to h_m and calculate the misclassification probability $P(T \Delta h_m)$.
    2. Calculate the side information sample $s(x_m)$ and update $\widetilde{h}_{m-1}$ to $\widetilde{h}_m$ by using both x_m and $s(x_m)$ and calculate the misclassification probability $P(T \Delta \widetilde{h}_m)$.

The above algorithm calculates the misclassification errors $P(T \Delta h_m)$ and $P(T \Delta \widetilde{h}_m)$ as a function of the sample size. Given $\epsilon > 0$ to find estimates for

$$P^m\{\bar{x} \in X^m \; ; \; h_m \text{ consistent with observations and } P(T \Delta h_m) > \epsilon\}$$

and

$$P^m\{\bar{x} \in X^m \; ; \; \widetilde{h}_m \text{ consistent with observations and } P(T \Delta \widetilde{h}_m) > \epsilon\}$$

we carry out N experiments as described above and for each m we calculate the fraction of experiment with misclassification error greater than ϵ.

For the simulations we chose $N = 16384$ and $M = 70$. Whenever analytically calculated learning error is available the simulated error agrees with it (see Figures 3.6 and 3.7). The simulation accuracy can be estimated by repeating the simulation with different seeds for the random number generator. For simulation with uniform distribution and the algorithm that chooses the smallest interval that is consistent with the observations, the maximum difference in six comparisons between different seeds was 0.096 without side information and 0.0132 with side information. Further, in all cases the maximum difference occured for sample 14 or smaller. An example of the simulation code is provided in Appendix B.

Density $2t$

When we calculated the learning rates with uniform distribution and side information, we observed the overlap phenomenon. Learning errors were devided in two cases: the first in which the overlap with the target and its side information translate was less than ϵ and the second in which the overlap was larger than ϵ or there was no overlap. In the second case the learning error was independent of the length of the target interval.

Now as expected the simulated results suggest that learning errors always depend on the location of the endpoints of the target interval. Figure 3.6 shows learning errors with side information for three different target intervals, together with the simulated error without side information and the analytically calculated error from Lemma 3.2.3.(Recall that without side information the learning error is independent of the density.) The simulated curve agrees with the analytical curve indicating the simulation accuracy.

Target $T = (0.05, 0.45)$ has the greatest benefit from the side information $s(x) = (x + 1/2) \bmod 1$ as it increases most the number of observations falling in $[0, 0.5]$. Similarly, the target $T = (0.4, 0.8)$ benefits from side information samples accumulating close to $1/2$. In the simulated examples the least advantage of side information occured for target $T = (1/2, b)$ such that $\int_{1/2}^{b} 2t\, dt > \epsilon$. We illustrate this by target $T = (0.5, 0.591608)$, where only a few side information translates are expected to fall inside the target. However, recall that the learning analysis is not this simple, one can not just look at the density of $s(x)$; as we discovered in the uniform distribution case that the overlap phenomena introduce other dependencies in the samples.

Normal Distribution

Next we simulate by drawing samples restricted to $[0, 1]$ from the normal distribution with mean $1/2$ and standard deviation 0.2, i.e., if a sample falls outside $[0, 1]$ we generate a new sample from the normal distribution, until a sample in $[0, 1]$ is obtained. Here the samples concentrate around $1/2$, whereas the side information translates $s(x) = (x + 1/2) \bmod 1$ concentrate around the endpoints 0 and 1. Thus the target $T = (0.1, 0.9)$ (see Figure 3.7) shows an enormous advantage of the side information and the target $T = (0.4, 0.6)$ gains only a small advantage as most samples produced by the side information are negative samples that are irrelevant for the algorithm.

The above setting can be carried even further. Let P_g have support on $[1/4, 3/4]$. Then for any target $T \subset [1/4, 3/4]$ the side information translates are negative samples and no advantage can be obtained. Interestingly, targets containing the support of the measure may benefit from the side information.

Averaging Algorithm with the Uniform Distribution

For the last illustration we change the learning algorithm. Now the estimate for the unknown target interval is as follows: the estimates for endpoints are the averages of the closest samples falling inside and outside the target interval. Let a_r denote the smallest positive sample and b_ℓ the largest positive sample, and let a_ℓ be the largest negative sample that is less than a_r and let b_r be the smallest negative sample that is greater than b_ℓ. So, a_ℓ is the closest sample "left" of unknown a and a_r is

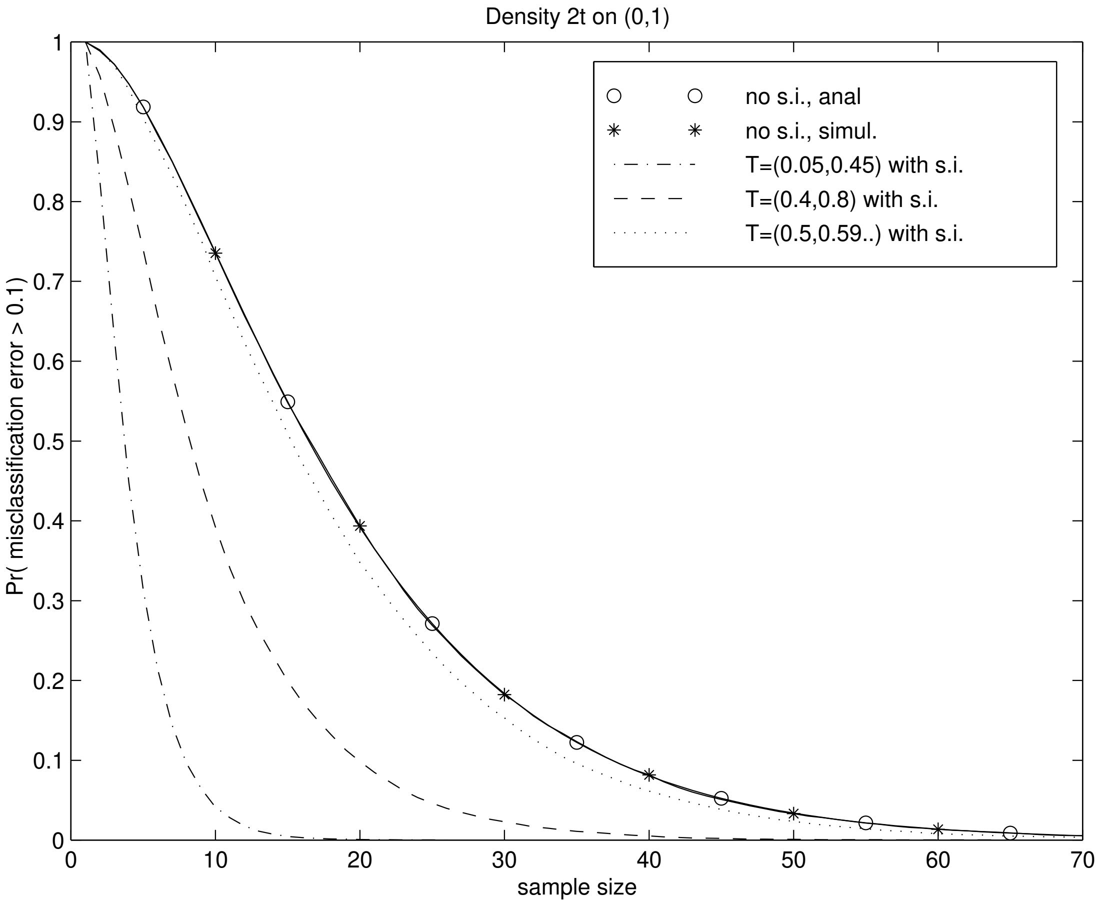

Figure 3.6: Simulation results with density $2t$. The y-axis is the empirical estimate for $P^m\{\bar{x} \in X^m \ ; \ h_m$ consistent with observations and $P(T\Delta h_m) > 0.1\}$ or $P^m\{\bar{x} \in X^m \ ; \ \widetilde{h}_m$ consistent with observations and $P(T\Delta h_m) > 0.1\}$ depending on the curve. Actually five learning rates have been plotted; the simulated and analytically calculated learning rates coincide in this scale.

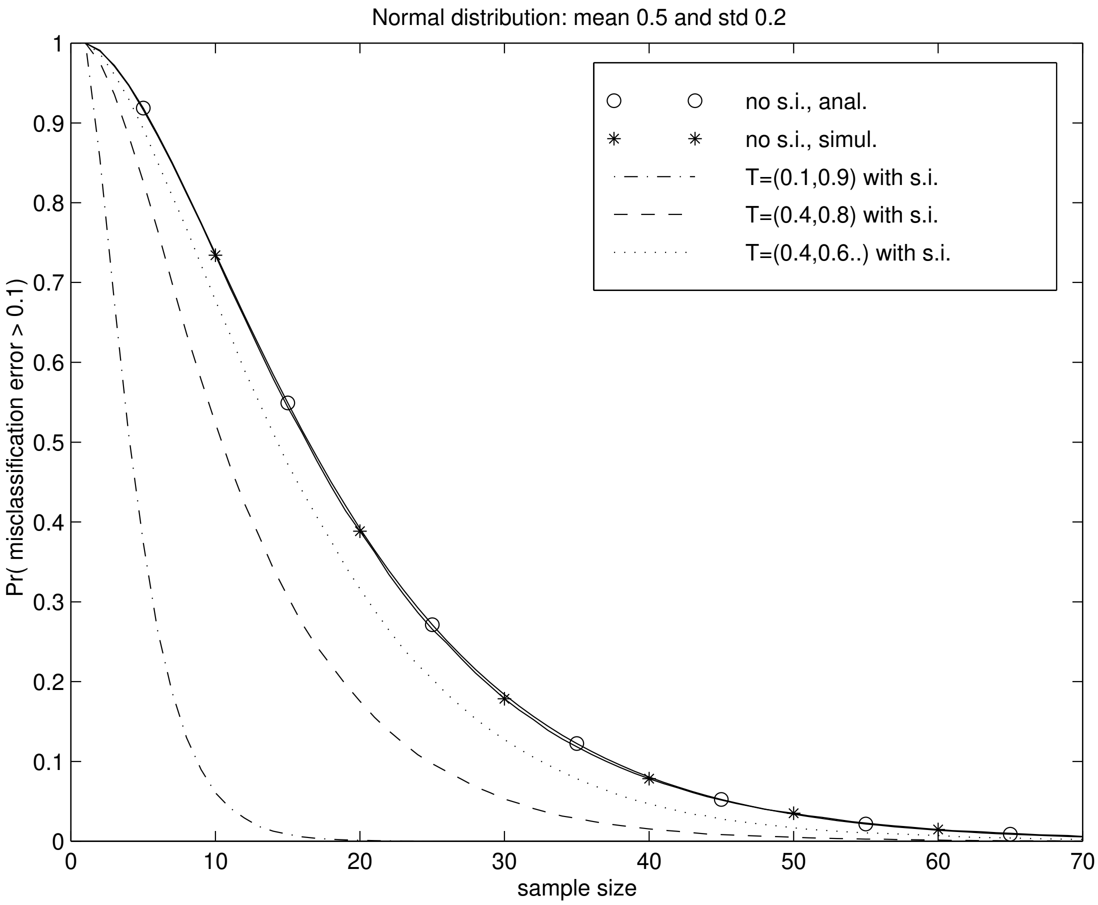

Figure 3.7: Simulation results by using the $(0,1)$ restricted normal distribution with mean 0.5 and standard deviation 0.2.

the closest sample on the "right" of a. For learning with side information similar variables are defined and they depend on the observations $x_1, \ldots$ together with the side information $s(x_1), \ldots$. We initialize the algorithm with $a_\ell = b_\ell = 0$ and $a_r = b_r = 1$, and similarly for learning with side information.

Recall that for the previous algorithm (approximation from inside) the learning error without side information was independent of the target intervals. Here it is no longer true. However, with or without side information the targets $T = (a, b)$ and $T = (1-b, 1-a)$ have the same learning rates due to the symmetry. Further, for the inside approximation the estimate with side information was always as good as or better than the estimate without it, regardless of the observations. In the averaging algorithm the side information estimate may be worse than the regular estimate. Figure 3.8 shows pairs of learning rates with and without side information for four targets. The curve with side information is always below the corresponding curve obtained without side information.

Figure 3.8 illustrates two types of behavior; one for targets $T = (0.4, 0.6)$ and $T = (0.15, 0.85)$ and one for targets $T = (0.2, 0.7)$ and $T = (0.2, 0.6)$. In the first type, for example, without side information and with only one sample, it is not possible to have an estimate with misclassification probability less than 0.1. This is because the endpoints are either close to 0 and 1 or both are close to 1/2. For $T = (0.2, 0.7)$ and $T = (0.2, 0.6)$ the probability of the above event (calculated analytically) is 0.2.

Remark 3.4.1. The averaging algorithm learns the target intervals a lot faster than the algorithm that approximates the target from inside. Figure 3.9 illustrates the target $T = (0.4, 0.6)$ and corresponding learning rates with both algorithms. Note that the averaging algorithm is an example of maximum margin algorithms that separate positive and negative observations with the maximum margin. Such algorithms usually perform well in learning applications.

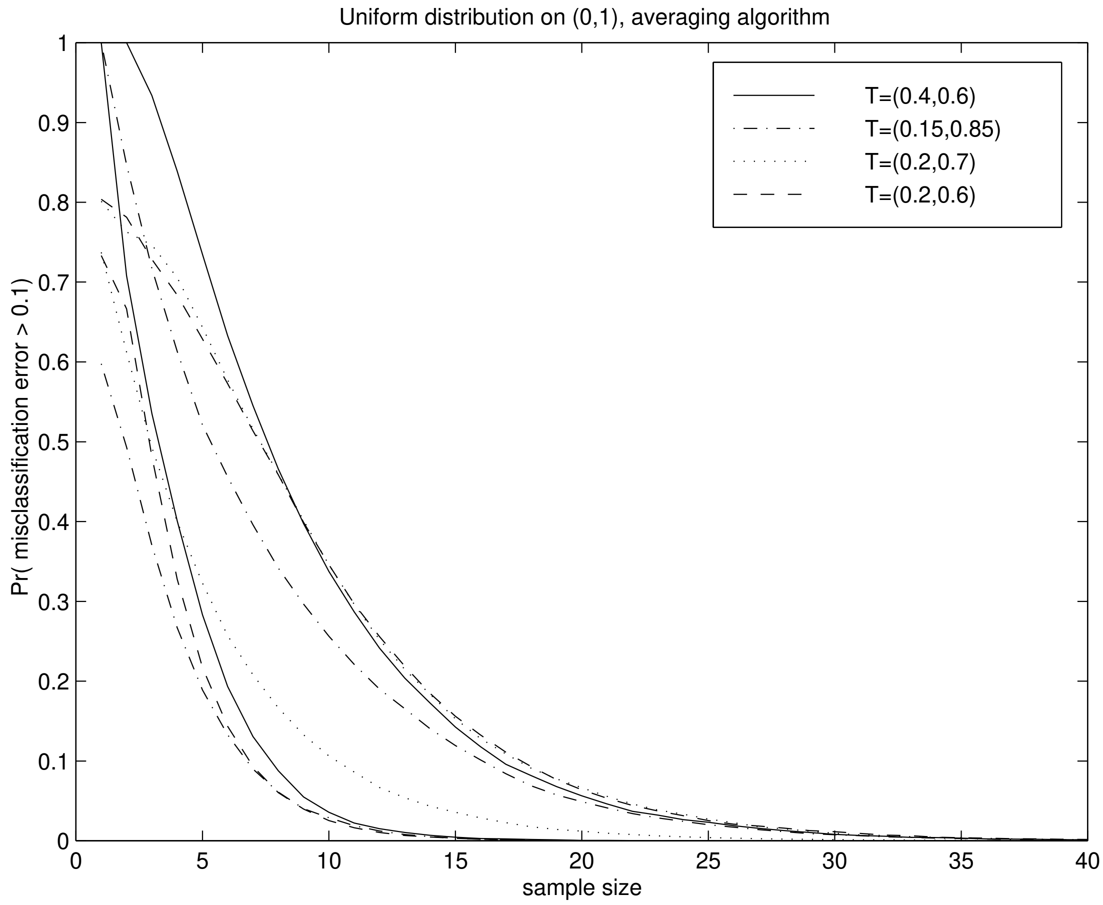

Figure 3.8: Simulation results with the averaging algorithm. A pair a learning rate curves is plotted for each target. In each pair, the upper curve corresponds to learning without side information.

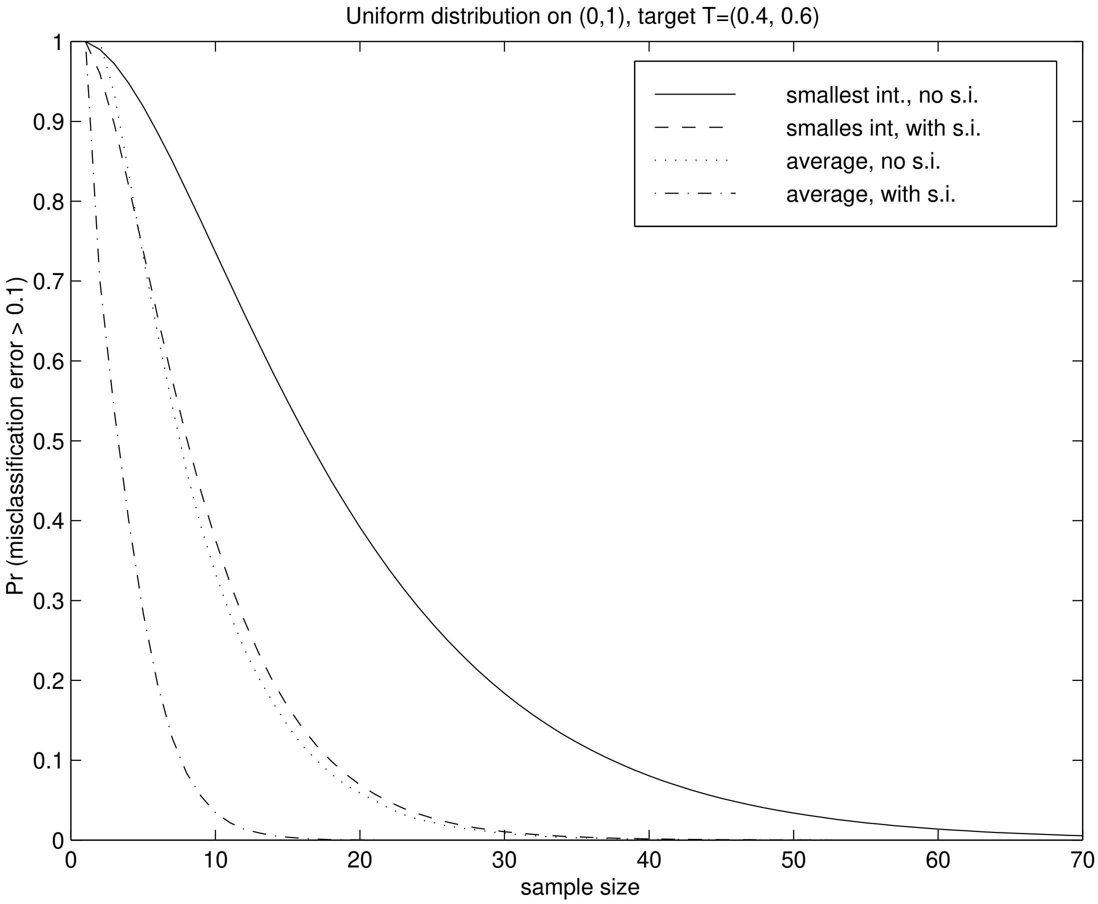

Figure 3.9: Comparison of the averaging algorithm and the smallest consistent interval algorithm with the target $T = (0.4, 0.6)$ and the uniform distribution.

Chapter 4

Side Information in the Learning Analysis

c:uniconv

In this chapter we discuss how side information can be incorporated in the learning theoretic analysis. Problems arise from two facts:

- The learning takes place with non-i.i.d. data.
- The spaces for training and classification are different.

In the presence of side information the available sample is $(x_1, s(x_1)), \ldots, (x_m, s(x_m))$ and an "oracle" will return both $\mathbf{1}_T(x_i)$ and $\mathbf{1}_T(s(x_i))$ for $i = 1, \ldots, m$. In this chapter our objective is to show improvement in the learning rate bounds of the classical learning theory when such side information is present. Let $A = h_m \Delta T$. There are two important factors for the improvement:

(i) the overlap of A and $s^{-1}(A)$

(ii) uniform upper and lower bounds on $\frac{P(s^{-1}(A))}{P(A)}$ for $A \in \mathcal{A}$.

The first factor we already saw in the exact rate computations of Chapter 3. The results in this chapter state improvement theorems for general learning spaces and consitent algorithms under assumptions concerning (i) and (ii).

First the extreme case of (ii) in which s is measure preserving is treated. With this assumption once s is given, it strongly restricts P. However, the techniques are easier to apply in this case, so it is presented first and then extended to almost measure preserving mappings.

If the side information mapping is measure preserving, side information samples can be treated in the same way as the original samples. In the interval problem we studied a specific learning problem and a fixed algorithm. By analyzing a measure preserving side information mapping for the exact calculations and by generalizing measures and side information mappings to study asympototic log learning rates, we saw that in some cases side information can lead to exponential improvement in the learning rate.

The learning rates for consistent algorithms are connected to the uniform convergence of empirical probabilities by the following inequality:

$$\begin{aligned} &P^m\{\bar{x} \in X^m \ ; \ \exists\, A \text{ consistent } , d_P(A) > \epsilon\} \\ &\qquad\qquad \leq P^m\{\bar{x} \in X^m \ ; \ \sup_{\substack{A\in\mathcal{A} \\ \sum \mathbf{1}_A(x_i)=0}} \left| \frac{1}{m}\sum_{i=1}^m \mathbf{1}_A(x_i) - P(A) \right| > \epsilon\}. \end{aligned}$$

Similarly, for learning problems in which zero empirical errors (i.e., consistency) can not be assumed, the error of the best classifier is bounded by the uniform convergence of empirical probabilities and we study (see Section 4.4)

$$P^m\{\bar{x} \in X^m \ ; \ \sup_{A\in\mathcal{A}} \left| \frac{1}{m}\sum_{i=1}^m \mathbf{1}_A(x_i) - P(A) \right| > \epsilon/2\}.$$

In Sections 4.1 and 4.2 our aim is to connect side information to the uniform convergence of empirical probabilities and to capture the exponential improvement from side information. First we analyze bounds of the form $C_1 e^{-c_2 m\epsilon^2}$ for the uniform convergence and we formulate the problem in the large deviations setting. In Section

4.2 we analyze with side information the bounds of the form $C_1 e^{-c_2 m\epsilon}$. In both sections we work with random variables $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 1, 2\}$, where $\mathbf{1}_A(x)$ is the characteristic functions of a set A. Depending on the probability of value 2, for example, exponentially faster convergence can be indicated. The special case $p_2 = 0$ can be derived by using existing techniques. However, we study the problem in general frameworks which allow to indicate the improvement when, say, p_2 is small.

We saw that the advantage of side information depends on the learning algorithm, distribution and the target concept. Recall that we apply the convergence bounds to sets $A = h\Delta T$, where h is an approximation of the unknown target T. Thus assumption on the distribution of $\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))$ is a synthesis of all components affecting the advantage of side information.

In the last section of this chapter we discuss a possible approach of introducing side information as a tool to obtain a new complexity measure. This is for cases in which side information samples and original samples are fundamentally different. This addresses the difference in the training and the classification spaces, while the actual learning takes place with i.i.d. samples.

4.1 Bounds of the Form $C_1 e^{-c_2 m\epsilon^2}$

s:sqbounds

In general, in function learning and in concept learning bounds of the form

$$P^m\big\{\bar{x} \in X^m \; ; \; \sup_{f\in\mathcal{C}} \Big|\frac{1}{m}\sum_{i=1}^{m} f(x_i) - E[f]\Big| \geq \epsilon\big\} \leq C_1 e^{-c_2 m\epsilon^2}$$

or

e:expterm

$$P^m\big\{\bar{x} \in X^m \; ; \; \sup_{A\in\mathcal{A}} \Big|\frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_A(x_i) - P(A)\Big| \geq \epsilon\big\} \leq C_1 e^{-c_2 m\epsilon^2}, \tag{4.4.1}$$

the exponential term comes from Hoeffding's inequality and the constant C_1 is usually a combinatorial term describing the richness of the classes $\mathcal{C}$ or $\mathcal{A}$.

Our aim is to study how side information can be incorporated into the concept learning problem in this setting and we indicate when side information can improve the exponential term in the bound (4.4.1). We carry out the analysis by assuming that the side information mapping is measure preserving, and later we show, see page 138, how that assumption can be relaxed for consistent learning algorithms.

4.1.1 Hoeffding's Inequality and Cramer's Theorem

In this section we review the basic tools for obtaining the exponential bounds. We begin by stating Hoeffding's inequality and we briefly sketch the idea of its proof, because we wish to establish a connection to the rate function approach we are going to follow. For that purpose we work with empirical probabilities rather than empirical sums.

Theorem 4.1.1 (Hoeffding's Inequality, [24]). *Let $X_1, \ldots, X_m$ be independent bounded random variables such that X_i falls in the interval $[a_i, b_i]$ with probability one. Denote their sum by $S_m = \sum_{i=1}^m X_i$. Then for any $\epsilon > 0$ we have*

$$\Pr\{S_m - E[S_m] \geq \epsilon\} \leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}$$

and

$$\Pr\{S_m - E[S_m] \leq -\epsilon\} \leq e^{-2\epsilon^2/\sum_{i=1}^m (b_i - a_i)^2}.$$

As a corollary,

$$\Pr\Big\{\big|1/m\sum_{i=1}^m \mathbf{1}_A(X_i) - P(A)\big| \geq \epsilon\Big\} \leq 2e^{-2m\epsilon^2}.$$

Sketch of proof. Hoeffding's inequality if proved by first showing that if $E[X] = 0$, $a \leq X \leq b$ and $s > 0$, then $E[e^{sX}] \leq e^{s^2(b-a)^2/8}$.

By using Markov's inequality, $\Pr\{X \geq \epsilon\} \leq E[X]/\epsilon$ for $\epsilon > 0$, we can write for $s > 0$, $\Pr\{X \geq \epsilon\} = \Pr\{e^{sX} \geq e^{s\epsilon}\} \leq e^{E[e^{sX}]}/e^{s\epsilon}$ and then $\Pr\{S_m - E[S_m] \geq \epsilon\} \leq e^{-s\epsilon} E\left[e^{s\sum(X_i - E[X_i])}\right]$.

Then for empirical probabilities

$$\begin{aligned}\Pr\Big\{1/m \sum_{i=1}^{m} \mathbf{1}_A(X_i) - P(A) \geq \epsilon\Big\} &= \Pr\Big\{S_m - mP(A) \geq m\epsilon\Big\} \\ &\leq e^{-sm\epsilon} E\left[e^{s\sum(X_i - E[X_i])}\right] = e^{-m\left(s\epsilon - \ln E\left[e^{s(X_i - E[X_i])}\right]\right)} \\ &\leq e^{-m \sup_{s>0}\left(s\epsilon - \ln E\left[e^{s(X_i - E[X_i])}\right]\right)}.\end{aligned}$$

By using the bound for $E[e^{sX}]$ and taking $s = 4\epsilon$ we see that

$$-m \sup_{s>0} \left(s\epsilon - \ln E\left[e^{s(X_i - E[X_i])}\right]\right) = -2m\epsilon^2.$$

The inequality for $\Pr\{1/m \sum_{i=1}^{m} \mathbf{1}_A(X_i) - P(A) \leq -\epsilon\}$ is proved analogously. □

Here the exponential rate was established by studying

$$\sup_{s>0} \left(s\epsilon - \ln E\left[e^{s(X_i - E[X_i])}\right]\right).$$

This will be seen as a rate function in the Large Deviations Principle that we state next.

Theorem 4.1.2 (Cramer's Theorem in $\mathbb{R}$ t:cramer [16]). *Let $\mu_m = 1/m \sum_{i=1}^{m} X_i$, where $X_1, \ldots, X_m$ are i.i.d. with probability law μ. Then $\{\mu_m\}$ satisfies the Large Deviation Principle with a convex rate function*

$$\Lambda^*(x) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \Lambda(\lambda)) = \sup_{\lambda \in \mathbb{R}} (\lambda x - \ln E[e^{\lambda X}]).$$

Further, for any closed set $F \subset \mathbb{R}$ and for all m

$$\mu_m(F) \leq 2e^{-m \inf_{x \in F} \Lambda^*(x)}.$$

For future reference we list some properties of $\Lambda(\lambda) = \ln E[e^{\lambda X}]$ and $\Lambda^*(x)$:

1. Λ is convex and Λ^* is a convex rate function.

2. For all $x > E[X]$, $\Lambda^*(x) = \sup_{\lambda \geq 0}(\lambda x - \Lambda(\lambda))$ and Λ^* is a nondecreasing function.

3. For all $x < E[X]$, $\Lambda^*(x) = \sup_{\lambda\leq 0}(\lambda x - \Lambda(\lambda))$ and Λ^* is a nonincreasing function.

4. When $E[X]$ is finite, $\Lambda^*(E[X]) = 0$ and always $\inf_{x\in\mathbb{R}} \Lambda^*(x) = 0$.

5.
$$\Lambda^*(x) = \inf\Big\{R(\nu \,\|\, \mu) \ ; \ \int_{\mathbb{R}} y\nu(dy) = x\Big\},$$
where $R(\nu \,\|\, \mu)$ is the relative entropy of ν with respect to μ.

An Application of Cramer's Theorem to Learning

Let $s : X \to X$ be measure preserving and let

$$Z_i^A = \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = \begin{cases} 0, & \text{with probability } p_0 \\ 1, & \text{with probability } p_1 \\ 2, & \text{with probability } p_2. \end{cases}$$

When $X_1, \ldots X_m$ are i.i.d. with probability law P, also $Z_1^A, \ldots Z_m^A$ are i.i.d. and we denote $S_m = Z_1^A + \cdots + Z_m^A$ and hence $E[S_m] = 2mP(A)$.

Now

$$\begin{aligned}\Pr\Big\{\big|\frac{1}{2m}S_m - P(A)\big| \geq \epsilon\Big\} &= \Pr\Big\{\frac{1}{m}S_m \in \big(-\infty, 2P(A) - 2\epsilon\big] \cup \big[2P(A) + 2\epsilon, \infty\big)\Big\} \\ &\leq 2\exp\big(-m \inf_{x\in F_\epsilon} \overset{\text{e:learn-cr}}{\Lambda^*_{\bar p}(x)}\big), \qquad (4.4.2)\end{aligned}$$

for all m and where

$$\begin{aligned}\bar p &= (p_0, p_1, p_2), \\ F_\epsilon &= \big(-\infty, 2P(A) - 2\epsilon\big] \cup \big[2P(A) + 2\epsilon, \infty\big) \text{ and} \\ \Lambda^*_{\bar p}(x) &= \sup_{\lambda\in\mathbb{R}}\big(\lambda x - \ln E[e^{\lambda Z}]\big).\end{aligned}$$

4.1.2 Bounding the Uniform Convergence, General Part

We formulate a general lemma for bounding the uniform convergence of empirical probabilities in the side information case. The proof follows standard procedure, and for the particular choice of parameter values we follow the proof of Devroye [17] for the first four steps. Reader who wants to skip the proof may resume reading from page 113. The essential idea is to use

$$\frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i))$$

to estimate $P(A)$.

l:genbnd

Lemma 4.1.3. *Let $\mathcal{A}$ be a permissible class (see Remark 4.1.9) and let $s : X \to X$ be measure preserving. Then*

$$\begin{aligned}
\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))-P(A)\right|\geq\epsilon\right\}& \\
&= \Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m} Z_i^A - P(A)\right| \geq \epsilon\right\}\\
&\leq 4E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})\exp\Big(-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*_{\bar{p}}(x)\Big)\right]\\
&\leq 4S(\mathcal{A},2m^2)E\left[\exp\Big(-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*_{\bar{p}}(x)\Big)\right],
\end{aligned}$$

where

$$\begin{aligned}
\epsilon' &= (m\epsilon-1)(m-1)/m^2,\\
\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2}) &= \left|\left\{\left(Z_1^A,\ldots,Z_{m^2}^A\right)\ ;\ A\in\mathcal{A}\right\}\right|,\\
S(\mathcal{A},n) &= \max_{x_1,\ldots,x_m}\left|\left\{\mathbf{1}_A(x_1),\ldots,\mathbf{1}_A(x_m)\ ;\ A\in\mathcal{A}\right\}\right|,
\end{aligned}$$

$\widetilde{\Lambda}^*_{\bar{p}}(x)$ is a rate function for random variable $\widetilde{Z}$ taken from population $z_1^A,\ldots,z_{m^2}^A$ with replacement and $F_{\epsilon'} = \left(-\infty,\widetilde{P}(A)-2\epsilon'\right]\cup\left[\widetilde{P}(A)+2\epsilon',\infty\right)$ with $\widetilde{P}(A) = 1/m^2\sum_{i=1}^{m^2} z_i^A$.

Proof. **Step 1: A ghost sample of size $m' = m^2 - m$**

Let $X_i \in X$ and let $s : X \to X$ be measure preserving. For a given $A \in \mathcal{A}$ we form $Z_i^A = 1_A(X_i) + 1_A(s(X_i)) = i$ with probability p_i for $i = 0, 1, 2$. We denote $T = X_1, \ldots, X_m$ and $V = X_{m+1}, \ldots, X_{m+m'}$, where $m' = m^2 - m$. Now $Z_1^A, \ldots, Z_{m^2}^A$ are i.i.d. random variables.

Let $\rho_A = |\widehat{P}_V(A) - \widehat{P}_T(A)|$, where $\widehat{P}_V(A)$ and $\widehat{P}_T(A)$ are the empirical probabilities

$$\widehat{P}_T(A) = \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = \frac{1}{2m} \sum_{i=1}^{m} Z_i^A \text{ and}$$
$$\widehat{P}_V(A) = \frac{1}{2m'} \sum_{i=1}^{m'} \mathbf{1}_A(X_{m+i}) + \mathbf{1}_A(s(X_{m+i})) = \frac{1}{2m'} \sum_{i=1}^{m'} Z_{m+i}^A$$

and denote $\rho = \sup_{A \in \mathcal{A}} \rho_A$ and $\sigma = \sup_{A \in \mathcal{A}} |P(A) - \widehat{P}_T(A)|$.

Throughout the proof we assume that

$$\sup_{A \in \mathcal{A}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A) \right|$$
$$\sup_{A \in \mathcal{A}} \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i))$$
$$\sup_{A \in \mathcal{A}} |\widehat{P}_T(A) - \widehat{P}_V(A)|$$

are measurable and $\sigma > \epsilon$ implies the there exists a random $A^* \in \mathcal{A}$ such that $\left| P(A^*) - \widehat{P}_T(A^*) \right| > \epsilon$. For more detailed discussion on measurability and selecting A^* see Remark 4.1.9.

In this step we link the uniform convergence of empirical probabilities in the m sample to the difference in empirical probabilites between the original sample and the ghost sample of size m'.

Proposition 4.1.4. *For all* $0 < \alpha < 1$ *and* $\epsilon > 0$,

$$\Pr\big\{\rho > (1 - \alpha)\epsilon\big\} \geq \Big(1 - \frac{1}{4\alpha^2 \epsilon^2 m'}\Big) \Pr\{\sigma > \epsilon\}.$$

Proof. Next we denote the probability measures in $(T, V) \in X^{m^2}$, $T \in X^m$ and $V \in X^{m'}$ by P, P_T and P_V, respectively.

The event $\{\sigma > \epsilon\}$ implies the existence of $A^* \in \mathcal{A}$ such that $|P(A^*) - \widehat{P}_T(A^*)| > \epsilon$ and on $\{\sigma > \epsilon\}$, the following holds: $\{|\widehat{P}_V(A^*) - P(A^*)| \leq \alpha\epsilon\} \subseteq \{\rho_{A^*} > (1-\alpha)\epsilon\} \subseteq \{\rho > (1-\alpha)\epsilon\}$.

Thus following Devroye [17]

$$\begin{aligned}
\Pr\{\rho > (1-\alpha)\epsilon\} &\geq \int_{X^{m^2}} \mathbf{1}_{\{\rho > (1-\alpha)\epsilon\}} dP \\
&= \int_{X^m} \int_{X^{m'}} \mathbf{1}_{\{\rho > (1-\alpha)\epsilon\}} dP_V dP_T \\
&\geq \int_{X^m} \mathbf{1}_{\{\sigma > \epsilon\}} \int_{X^{m'}} \mathbf{1}_{\{\rho > (1-\alpha)\epsilon\}} dP_V dP_T \\
&\geq P_T\{\sigma > \epsilon\} \inf_{A \in \mathcal{A}} P_V\big\{|\widehat{P}_V(A) - P(A)| \leq \alpha\epsilon\big\} \\
&= P_T\{\sigma > \epsilon\} \inf_{A \in \mathcal{A}} \Big(1 - P_V\big\{|\widehat{P}_V(A) - P(A)| > \alpha\epsilon\big\}\Big) \\
&\geq P_T\{\sigma > \epsilon\} \Big(1 - \frac{1}{4\alpha^2\epsilon^2 m'}\Big) \quad \text{by Remarks 4.1.5 and 4.1.6.}
\end{aligned}$$

Remark $\overset{\text{r:step1a}}{4.1.5}$. If $Z \in \{0, 1, 2\}$ then $\mathrm{Var}(Z) \leq 1$.

Proof. $\mathrm{Var}(Z) = E\big[(Z - E[Z])^2\big] \leq \frac{2-E[Z]}{2} E[Z]^2 + \frac{E[Z]}{2}(2 - E[Z])^2 = E[Z](2 - E[Z]) \leq 1$, where the first inequality follows from the convexity of x^2. ☐

Remark $\overset{\text{r:step1b}}{4.1.6}$.

$$\begin{aligned}
P_V\big\{|\widehat{P}_V(A) - P(A)| > \alpha\epsilon\big\} &= P_V\Big\{\Big|\frac{1}{2m'}\sum_{i=1}^{m'} Z^A_{m+i} - P(A)\Big| > \alpha\epsilon\Big\} \\
&= P_V\Big\{\Big|\frac{1}{m'}\sum_{i=1}^{m'} Z^A_{m+i} - 2P(A)\Big| > 2\alpha\epsilon\Big\} \leq \frac{\mathrm{Var}\, Z^A_{m+i}}{m'(2\alpha\epsilon)^2} \leq \frac{1}{4\alpha^2\epsilon^2 m'},
\end{aligned}$$

where the first inequality follows from Chebychev's inequality and the last inequality uses the Remark 4.1.5.

☐

Step 2: Permutations

Let (T_i, V_i) denote one of the possible $m^2!$ permutations of (T, V) and introduce $\rho_A(i) = |\widehat{P}_{V_i}(A) - \widehat{P}_{T_i}(A)|$ and all other notations similarly.

Sets A and B are equivalent for (T, V) if

$$A \cap \{X_1, s(X_1), \ldots, X_{m^2}, s(X_{m^2})\} = B \cap \{X_1, s(X_1), \ldots, X_{m^2}, s(X_{m^2})\}.$$

When A and B are equivalent for (T, V), also $\widehat{P}_{V_i}(A) = \widehat{P}_{V_i}(B)$ and $\widehat{P}_{T_i}(A) = \widehat{P}_{T_i}(B)$ for all $i = 1, \ldots, m^2!$.

We bound the fraction of permutations such that $\rho(i) > (1-\alpha)\epsilon$:

$$\begin{aligned}\frac{1}{m^2!}\sum_{i=1}^{m^2!} \mathbf{1}_{\{\rho(i)>(1-\alpha)\epsilon\}} &= \frac{1}{m^2!}\sum_{i=1}^{m^2!} \sup_{A\in\mathcal{A}} \mathbf{1}_{\{\rho_A(i)>(1-\alpha)\epsilon\}} \\ &\le \frac{1}{m^2!}\sum_{i=1}^{m^2!} \sum_{A\in\mathcal{A}(T,V)} \mathbf{1}_{\{\rho_A(i)>(1-\alpha)\epsilon\}} = \sum_{A\in\mathcal{A}(T,V)} \frac{1}{m^2!}\sum_{i=1}^{m^2!} \mathbf{1}_{\{\rho_A(i)>(1-\alpha)\epsilon\}},\end{aligned}$$

where $\mathcal{A}(T, V) \subseteq \mathcal{A}$ is a subclass such that

(i) $A, B \in \mathcal{A}(T, V) \Rightarrow A$ and B are not equivalent for (T, V)

(ii) for every $A \in \mathcal{A}$ there exists $B \in \mathcal{A}(T, V)$ that is equivalent to A for (T, V).

Step 3: Conditioning

Fix $X_1 = x_1, \ldots, X_{m^2} = x_{m^2}$ and the corresponding sequence $z_1^A, \ldots, z_{m^2}^A$. For notational convinience we drop the superscript A for a moment.

Then

$$\begin{aligned}&\Pr\left\{ |\frac{1}{2m}\sum_{i=1}^{m} z_i - \frac{1}{2m'}\sum_{i=1}^{m'} z_{m+i}| > (1-\alpha)\epsilon \right\} \\ &\qquad = \Pr\left\{ \left| \frac{1}{2m}\sum_{i=1}^{m} z_i - \frac{1}{2m'}\left(2m^2\widehat{P}_{(T,V)}(A) - \sum_{i=1}^{m} z_i\right) \right| > (1-\alpha)\epsilon \right\},\end{aligned}$$

where

$$\widehat{P}_{(T,V)}(A) = \frac{1}{2(m+m')}\left(\sum_{i=1}^{m} z_i + \sum_{i=1}^{m'} z_{m+i}\right).$$

Simplifying further we get

$$\begin{aligned}
\Pr\left\{\left|\frac{1}{2m}\sum_{i=1}^{m} z_i - \frac{1}{2m'}\left(2m^2\widehat{P}_{(T,V)}(A) - \sum_{i=1}^{m} z_i\right)\right| > (1-\alpha)\epsilon\right\}& \\
&= \Pr\left\{\left|\frac{m^2}{m'}\frac{1}{2m}\sum_{i=1}^{m} z_i - \frac{m^2}{m'}\widehat{P}_{(T,V)}(A)\right| > (1-\alpha)\epsilon\right\} \\
&= \Pr\left\{\left|\frac{1}{2m}\sum_{i=1}^{m} z_i - \widehat{P}_{(T,V)}(A)\right| > (1-\alpha)\epsilon\frac{m'}{m^2}\right\}
\end{aligned}$$

Let us denote $\epsilon' = (1-\alpha)\epsilon m'/m^2$.

Lemma 4.1.7 (Hoeffding's lemma [24]). *Let C be a population consisting of m^2 values $c_1, \ldots, c_{m^2}$. Let $X_1, \ldots, X_m$ denote a random sample from C without replacement and let $Y_1, \ldots, Y_m$ be a sample with replacement (and hence i.i.d.) from C.*

If f is continuous and convex then

$$E[f(\sum_{i=1}^{m} X_i)] \le E[f(\sum_{i=1}^{m} Y_i)].$$

We apply Hoeffding's lemma to $f(x) = e^{sx}$ in the Cramer's theorem. The notation $\overline{\Pr}$ refers to the probability over random selections from a fixed sample $z_1, \ldots, z_{m^2}$. Thus

$$\overline{\Pr}\left\{\left|\frac{1}{2m}\sum_{i=1}^{m} z_i - \widehat{P}_{(T,V)}(A)\right| > \epsilon'\right\} \le 2e^{-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*(x)},$$

where

$$\begin{aligned}
\widetilde{\Lambda}^*(x) &= \sup_{\lambda\in\mathbb{R}}\left(\lambda x - \ln E[e^{\lambda\widetilde{Z}}]\right), \\
F_{\epsilon'} &= \left(-\infty, 2\widehat{P}_{(T,V)}(A) - 2\epsilon'\right] \cup \left[2\widehat{P}_{(T,V)}(A) + 2\epsilon', \infty\right),
\end{aligned}$$

and $\widetilde{Z}$ is a random variable taken from population $z_1, \ldots, z_{m^2}$ with replacement. Further $\widetilde{Z} = j$ with probability $\hat{p}_j$ for $j = 0, 1, 2$, where $\hat{p}_j = 1/(m^2)\sum_{i=1}^{m^2}\mathbf{1}_{\{z_i=j\}}$ for $j = 0, 1, 2$.

Step 4: Collecting bounds

Combining Steps 2 and 3, and taking an expectation gives

$$\Pr\{\rho > (1-\alpha)\epsilon\} \le 2E\left[\widetilde{N}_{\mathcal{A}}\left(X_1,\ldots,X_{m^2}\right) e^{-m\inf_{x\in F_{\epsilon'}} \widetilde{\Lambda}^*(x)}\right],$$

where

$$\widetilde{N}_{\mathcal{A}}\left(X_1,\ldots,X_{m^2}\right) = |\{\left(Z_1^A,\ldots,Z_{m^2}^A\right)\ ;\ A\in\mathcal{A}\}|.$$

Together with Step 1 this gives us

$$\begin{aligned}\Pr\{\sigma > \epsilon\} &\le 2E\left[\widetilde{N}_{\mathcal{A}}\left(X_1,\ldots,X_{m^2}\right) e^{-m\inf_{x\in F_{\epsilon'}} \widetilde{\Lambda}^*(x)}\right] \frac{1}{1-\frac{1}{4m'\alpha^2\epsilon^2}} \\ &\le 2\frac{1}{1-\gamma^2/2} E\left[\widetilde{N}_{\mathcal{A}}\left(X_1,\ldots,X_{m^2}\right) e^{-m\inf_{x\in F_{\epsilon'}} \widetilde{\Lambda}^*(x)}\right],\end{aligned}$$

when $\gamma = 1/(m\epsilon\alpha)$, $m \ge 2$ and $0 < \gamma \le \sqrt{2}$.

By taking $\gamma = 1$, we get

$$\Pr\{\sigma > \epsilon\} \le 4E\left[\widetilde{N}_{\mathcal{A}}\left(X_1,\ldots,X_{m^2}\right) e^{-m\inf_{x\in F_{\epsilon'}} \widetilde{\Lambda}^*(x)}\right],$$

where $\epsilon' = (m\epsilon - 1)(m-1)/m^2$.

Step 5: Bounding $\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_m)$

Recall that $Z_i^A = \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0,1,2\}$ and for a moment let $f_A(x) = \mathbf{1}_A(x) + \mathbf{1}_A(s(x))$.

p:comb-term

Proposition 4.1.8.

$$\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_m) \le \widetilde{S}(\mathcal{A},m) \le S(\mathcal{A},2m),$$

where

$$\begin{aligned}\widetilde{S}(\mathcal{A},m) &= \max_{(x_1,\ldots,x_m)} |\{(f_A(x_1),\ldots,f_A(x_m))\ ;\ A\in\mathcal{A}\}|, \\ S(\mathcal{A},m) &= \max_{(x_1,\ldots,x_m)} |\{(\mathbf{1}_A(x_1),\ldots,\mathbf{1}_A(x_m))\ ;\ A\in\mathcal{A}\}|\end{aligned}$$

Proof. Fix $(x_1, \ldots, x_m)$. Each $A \in \mathcal{A}$ gives rise to

$$(\mathbf{1}_A(x_1), \mathbf{1}_A(s(x_1)), \ldots, \mathbf{1}_A(x_m), \mathbf{1}_A(s(x_m)))$$

and if

$$(f_A(x_1), \ldots, f_A(x_m)) \neq (f_{A'}(x_1), \ldots, f_{A'}(x_m))$$

for some $A, A' \in \mathcal{A}$, then also

$$\begin{aligned}(\mathbf{1}_A(x_1), \mathbf{1}_A(s(x_1)), \ldots, \mathbf{1}_A(x_m), \mathbf{1}_A(s(x_m)))& \\ \neq (\mathbf{1}_{A'}(x_1), \mathbf{1}_{A'}(s(x_1)), \ldots, \mathbf{1}_{A'}(x_m), \mathbf{1}_{A'}(s(x_m)))&.\end{aligned}$$

Thus

$$\begin{aligned}&\left|\left\{\left(f_A(x_1), \ldots, f_A(x_m)\right)\ ;\ A \in \mathcal{A}\right\}\right| \\ &\quad \leq \left|\left\{\left(\mathbf{1}_A(x_1), \mathbf{1}_A(s(x_1)), \ldots, \mathbf{1}_A(x_m), \mathbf{1}_A(s(x_m))\right)\ ;\ A \in \mathcal{A}\right\}\right| \\ &\quad \leq \max_{(x_1, x_2, \ldots, x_{2m-1}, x_{2m})} \left|\left\{\left(\mathbf{1}_A(x_1), \mathbf{1}_A(x_2), \ldots, \mathbf{1}_A(x_{2m-1}), \mathbf{1}_A(x_{2m})\right)\ ;\ A \in \mathcal{A}\right\}\right| \\ &\quad = S(\mathcal{A}, 2m).\end{aligned}$$

Hence $\widetilde{S}(\mathcal{A}, m) \leq S(\mathcal{A}, 2m)$. □

Combining Step 5 with m^2 in place of m and Step 4 yields the result. □

r:meas

Remark 4.1.9. The 5 step process above assumed that $\sup_{A\in\mathcal{A}} |1/(2m) \sum_{i=1}^m Z_i^A - P(A)|$, $\sup_{A\in\mathcal{A}} 1/(2m) \sum_{i=1}^m Z_i^A$ and $\sup_{A\in\mathcal{A}} |1/(2m) \sum_{i=1}^m Z_i^A - 1/(2m') \sum_{i=1}^{m'} Z_{m+i}^A|$ were measurable. The next discussion is mostly based on an Appendix in [30], which in turn relies on [15] for proofs.

Let spaces X, Y be homeomorphic to a complete separable metric space and its subspace and let $N = \mathbb{R} \setminus \mathbb{Q}$ be the set of irrational numbers of $\mathbb{R}$ with metric $d(x, y) = |x - y|$. The class of analytic sets is an extension of Borel sets: a subset $S \subseteq X$ is called analytic if it satisfies any of the following equivalent conditions: (i)

S is a continuous image of N, (ii) S is a continuous image of a Borel set in X, (iii) S is the projection of a closed set in a product space $X \times N$, (iv) S is the projection of a Borel set in $X \times Y$. We denote the class of analytic sets of X by $A(X)$.

A more general definition for analytic sets is given in [15]. Let F be a set. A *paving* of F is a family of subsets of F which contains the empty set; the pair $(F, \mathcal{F})$ consisting of a set F and a paving $\mathcal{F}$ is called a *paved* set. A subset A of F is called $\mathcal{F}$-analytic if there exists an auxiliary compact metrizable space E and a subset $B \subset E \times F$ belonging to $(\mathcal{K}(E) \times \mathcal{F})_{\sigma\delta}$, such that A is the projection of B onto F. In the above definition, $\mathcal{K}(E)$ denotes the compact subsets of E and $()_{\sigma\delta}$ denotes the closure under finite unions and arbitrary intersections.

Without loss of generality, we assume that there exists a class $\mathcal{H}$ that is indexed by a parameter, i.e., $\mathcal{H} = \{h(\cdot, t) \ ; \ t \in T\}$. (We can take $T = \mathcal{H}$). Assume that T is a separable metric space and denote the Borel σ-field induced by the metric as $\mathcal{B}(T)$.

d:permis

Definition 6. $\mathcal{H}$ is called *permissible* if it can be indexed by a T in such a way that

(i) the function $h(\cdot, \cdot)$ is $\mathcal{S} \times \mathcal{B}(T)$-measurable as a function from $S \times T$ into the real line, where $\mathcal{S}$ is the σ-field of S.

(ii) T is an analytic subset of a compact metric space $\bar{T}$. (This condition is called as Souslin measurable space by many authors.)

Let $\xi_1, \xi_2, \ldots$ be measurable maps from a complete probability space $(\Omega, \mathcal{E}, \mathbb{P})$ into a set S equipped with a σ-field $\mathcal{S}$. Assume that a class $\mathcal{H}$ of $\mathcal{S}/\mathcal{B}(\mathbb{R})$-measurable functions is permissible. Next we list some properties of analytic sets. For any measurable space $(X, \mathcal{X})$:

(a) $A(X \times T)$ contains the product σ-field $\mathcal{X} \times \mathcal{B}(T)$.

(b) for each L in $A(X \times T)$, and in particular for each $\mathcal{X} \times \mathcal{B}(T)$-measurable set, the projection $\Pi_X L$ of L to X is in $A(X)$.

(c) for each C in $A(X)$ and each $\mathcal{E}/\mathcal{X}$-measurable map η from $(\Omega, \mathcal{E}, \mathbb{P})$ into X, the set $\{\eta \in C\}$ is an analytic subset of Ω; hence $\{\eta \in C\}$ belongs to $\mathcal{E}$ because $(\Omega, \mathcal{E}, \mathbb{P})$ is complete. (If $\mathcal{E}$ is complete, then $A(\Omega) = \mathcal{E}$.)

We show next that if $\mathcal{A}$ is permissible then $\sup_{A\in\mathcal{A}} |1/(2m) \sum_{i=1}^m \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A)| = \sup_{A\in\mathcal{A}} |1/(2m) \sum_{i=1}^m Z^A(X_i) - P(A)|$ is measurable, the two other suprema are done similarly. Let $h(\xi_i(\omega)) = h(s_i, t) = Z^t(s_i)$ and assume that $\mathcal{H} = \{h(\cdot, t) \ ; \ t \in T\}$ is permissible. Let $\eta = (\xi_1, \ldots, \xi_m)$, $s = (s_1, \ldots, s_m)$ and define $g(s,t) = |1/(2m) \sum_{i=1}^m h(s_i, t) - P(h_t)|$, which is a $\mathcal{S}^m \times \mathcal{B}(T)$-measurable real function on $S^m \times T$. Let $G(m) = \sup_{t\in T} g(m, t)$.

By (a) $A(S^m \times T)$ contains the set $L_\alpha = \{(m,t) \ ; \ g(m,t) > \alpha\}$ and the projection of L_α onto S^m is an analytic set by (b). The projection consists of all m such that $G(m) > \alpha$, thus $\{G > \alpha\}$ is in $A(S^m)$ for each real α. As $(\xi_1, \ldots, \xi_m)$ is a measurable map from a complete probability space $(\Omega, \mathcal{E}, \mathbb{P})$ into S^m, by (c), we can conclude that $\{\omega \ ; \ G(\eta(\omega)) > \alpha\}$ is $\mathcal{E}$-measurable. Thus $\sup_{t\in T} g(\eta(\omega), t)$ is a $\mathcal{E}$-measurable function of ω, i.e., $\sup_{t\in T} |1/(2m) \sum_{i=1}^m h(s_i, t) - P(h_t)|$ is measurable.

Typical classes parametrized by Borel subsets of an euclidean space are permissible. We quote a theorem from [15, p. 43]:

Theorem 4.1.10 ([15]). *$A(\mathcal{H})$ contains the σ-field $\sigma(\mathcal{H})$ generated by $\mathcal{H}$ if and only if the complement of every element of $\mathcal{H}$ is analytic.*

We also assumed in the proof that $\sup_{A\in\mathcal{A}} \left| \widehat{P}_T(A) - P(A) \right| > \epsilon$ implies that there exists a random A^* with the property $\left| \widehat{P}_T(A^*) - P(A^*) \right| > \epsilon$. This can be proved by using the cross-section theorem:

Theorem 4.1.11 ([15, 30]). *Under the requirement (ii) of Definition 6 and for any complete probability space $(M, \mathcal{M}, \mu)$ the following holds: If H belongs to $A(M \times T)$*

there exists a measurable map g from M into $\bar{T} = T \cup \{\infty\}$ such that: $(m, g(m))$ belongs to H whenever $g(m) \neq \infty$ and $g(m) \neq \infty$ for μ almost all m in the projection $\Pi_M H$. We call g as measurable cross-section for H.

4.1.3 Bounding the Uniform Convergence, Exponential Term

s:sqb-expterm

So far we have just assumed that the side information mapping $s(x)$ is measure preserving. However, more has to be assumed for improved rates. For example, $s(x) = x$ is measure preserving but it provides no new information, indeed, in this case $\left|\frac{1}{2m}\sum_{i=1}^m \mathbf{1}_A(x) + \mathbf{1}_A(s(x)) - P(A)\right| = |1/m \sum_{i=1}^m \mathbf{1}_A(x) - P(A)|$, which is just the quantity obtained with no side information.

Each $A \in \mathcal{A}$ gives rise to $\bar{p}_A = (p_0, p_1, p_2)$ such that $Z^A(x) = \mathbf{1}_A(x) + \mathbf{1}_A(s(x)) = i$ with probability p_i for $i = 0, 1, 2$ and let $\mathbb{P}_{\mathcal{A}} = \{\bar{p}_A \,;\ A \in \mathcal{A}\}$.

If $(p, 0, 1-p) \in \mathbb{P}_{\mathcal{A}}$ for all $p \in [0, 1]$ no advantage of side information can be obtained.

We illustrate cases in which side information gives exponential improvement in the learning rate bound.

Case 1: "No overlap"

Assume that $\bar{p}_A = (p, 1-p, 0)$ for each $A \in \mathcal{A}$. This situation has been studied in the interval learning problem when the target $T \subset (0, 1/2)$ and $s(x) = (x + 1/2)$ mod 1.

t:sqb-p2-0

Theorem 4.1.12. *Let $\mathcal{A}$ be a permissible class. If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 1\}$ for* *all i and $s : X \to X$ is measure preserving then*

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))-P(A)\right|\geq\epsilon\right\}$$
$$\leq 4e^{16\epsilon^2+16\epsilon}E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})\right]e^{-8m\epsilon^2}$$
$$\leq 4e^{16\epsilon^2+16\epsilon}S(\mathcal{A},2m^2)e^{-8m\epsilon^2}.$$

Proof. In this case the $\hat{p}=(\hat{p}_0,\hat{p}_1,\hat{p}_2)$, at the end of the Step 3 in the proof of Lemma 4.1.3, satisfies $\hat{p}\in\mathbb{P}_{\mathcal{A}}$ and

$$E\left[e^{-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*_{\hat{p}}(x)}\right]\leq e^{-m\inf_{\tilde{p}\in\mathbb{P}_{\mathcal{A}}}\inf_{x\in F_{\epsilon'}}\Lambda^*_{\tilde{p}}(x)},$$

where $\epsilon'=(m\epsilon-1)(m-1)/m^2$.

When calculating the rate we use the property that large deviations rate function can be expressed as a relative entropy. Here with notation $0\ln 0:=0$ and $0\ln 0/0:=0$,

$$\Lambda^*_{\tilde{p}}(x)=\Lambda^*_p(x)=\begin{cases}x\ln\frac{x}{p}+(1-x)\ln\frac{1-x}{1-p}, & \text{if } x\in[0,1],\\ \infty, & \text{otherwise,}\end{cases}$$

and we can take $F_{\epsilon'}=(-\infty,p-2\epsilon']\cup[p+2\epsilon',\infty)$.

By using the convexity of $\Lambda^*_p(x)$,

$$\inf_{x\in F_{\epsilon'}}\Lambda^*_p(x)=\min\{\Lambda^*_p(p+2\epsilon'),\Lambda^*_p(p-2\epsilon')\},$$

and the properties of $\Lambda^*_p(x)$ (namely, $\Lambda^*_p(p)=0$ and $\inf_{x\in\mathbb{R}}\Lambda^*_p(x)=0$) give us

$$\Lambda^*_p(p+2\epsilon')=\Lambda^*_p(p)+\Lambda^{*\prime}_p(p)2\epsilon'+(1/2)\Lambda^{*\prime\prime}_p(\theta)(2\epsilon')^2=(1/2)\Lambda^{*\prime\prime}_p(\theta)(2\epsilon')^2,$$

where $\theta\in[p,p+2\epsilon']$, and similarly for $\Lambda^*_p(p-2\epsilon')$.

Now

$$\Lambda^{*\prime\prime}_p(x)=\frac{1}{(1-x)x}\geq 4 \text{ for } x\in(0,1)$$

and $\inf_{x\in F_{\epsilon'}}\Lambda^*_p(x)\geq(1/2)4(2\epsilon')^2=8\epsilon'^2$. As $e^{-8m\epsilon'^2}\leq e^{16\epsilon^2+16\epsilon}e^{-8m\epsilon^2}$, this together with the Lemma 4.1.3 implies the result. □

Note that this result can be obtained by using Hoeffding's inequality as well.

In comparison to the best exponent by Devroye [17]

e:devbnd
$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|\geq\epsilon\right\}\leq 4e^{4\epsilon+4\epsilon^2}S(\mathcal{A},m^2)e^{-2m\epsilon^2} \tag{4.4.3}$$

we see that we have been able to improve the exponent by the factor of 4. Although the combinatorial term is worse, that term grows only polynomially in the sample size and the improvement in the exponential term gives a better bound unless m is small.

Case 2: "Small overlap"

We assume that the overlap of A and $s^{-1}(A)$ is small and we consider

$$\mathbb{P}_{\mathcal{A}}=\{(p_0,p_1,p_2)\ ;\ p_i\geq 0,\sum p_i=1,p_2\leq\beta\}.$$

The rate function for $p_i>0$, $p_0+p_1+p_2=1$ and $0\leq x\leq 2$ is given by

$$\begin{aligned}
&\Lambda^*_{\bar{p}}(x)\\
&=\Lambda^*_{p_1,p_2}(x)=\inf_{q_0,q_1,q_2}\{q_0\ln\frac{q_0}{p_0}+q_1\ln\frac{q_1}{p_1}+q_2\ln\frac{q_2}{p_2}\ ;\ q_1+2q_2=x,q_0+q_1+q_2=1\}\\
&\qquad=\inf_{\substack{\max(0,x-1)\leq q\\ \leq\min(1,x/2)}}\{(1-x+q)\ln\frac{1-x+q}{1-p_1-p_2}+(x-2q)\ln\frac{x-2q}{p_1}+q\ln\frac{q}{p_2}\}.
\end{aligned}$$

We begin by showing that $\mathbb{P}_{\mathcal{A}}$ has to be relaxed to allow larger values for p_2 and the convergence bound can be split into two parts: one occuring with high probability in which we have control over p_2 and one with small probability, where we apply existing bounding techniques.

t:sip2
Theorem 4.1.13. *Assume that $\mathcal{A}$ is permissible. If $s : X\to X$ is measure pre-*

serving and $\mathbb{P}_{\mathcal{A}} = \{(p_0, p_1, p_2) \;;\; p_0 + p_1 + p_2 = 1, p_2 \leq \beta\}$ *then*

$$\begin{aligned}\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A)\right| > \epsilon\right\}\\
&\leq 4E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})e^{-m\inf_{p\in\mathbb{P}_{\mathcal{A}}}\inf_{x\in F_{\epsilon'}}\Lambda^*_{\hat{p}}(x)}\right]\\
&\leq 4S(\mathcal{A},2m^2)\left(\left(1-C(d,m^2)\right)e^{-m\inf_{p\in\mathbb{P}_{\mathcal{A},d}}\inf_{x\in F_{\epsilon'}}\Lambda^*_{\hat{p}}(x)}\right.\\
&\qquad\qquad \left. + C(d,m^2)e^{-2m\epsilon^2+4\epsilon+4\epsilon^2}\right),\end{aligned}$$

where $\mathbb{P}_{\mathcal{A},d} = \{(p_0, p_1, p_2) \;;\; p_0 + p_1 + p_2 = 1, p_2 \leq \beta + d\}$ *and* $C(d,m) = 4e^{4d^2+4d}S(\mathcal{A}, 2m^2)e^{-2d^2m}$.

We want to emphasize that the sample size in the "correction term" $C(d,m^2) = 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}$ is m^2 and hence that term becomes negligible very quickly and the exponential term from the rate function dominates.

Proof. The result follows from Lemma 4.1.3 after solving a technical difficulty: $(\hat{p}_0, \hat{p}_1, \hat{p}_2)$ obtained in Step 3 of Lemma 4.1.3 may not be in $\mathbb{P}_{\mathcal{A}}$ although (p_0, p_1, p_2) (or A) that generated the sequence $z_1, \ldots, z_{m^2}$ is in $\mathbb{P}_{\mathcal{A}}$. However, note that $\hat{p}_2 \to p_2$ in probability and we get a bound:

p:corrterm

Proposition 4.1.14.

$$\Pr\{\sup_{A\in\mathcal{A}}\hat{p}_2 - p_2 > d\} \leq 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}.$$

Proof. Recall the result by Devroye [17]:

$$\Pr\left\{\sup_{\hat{A}\in\hat{\mathcal{A}}}\left|\frac{1}{n}\sum_{i=1}^{n}\mathbf{1}_{\hat{A}}(X_i) - P(\hat{A})\right| > \epsilon\right\} \leq 4e^{4\epsilon^2+4\epsilon}S(\hat{\mathcal{A}},n^2)e^{-2\epsilon^2n^2}.$$

Now $\hat{p}_2 = 1/m^2\sum_{i=1}^{m^2}\mathbf{1}_{\{Z_i^A=2\}}$ and define $W_i^A = W^A(X_i) = 1$, if $Z_i^A = 2$ and 0 otherwise. This give rise to a new collection of sets (namely $\hat{A} = A \cap s^{-1}(A)$) as A ranges through $\mathcal{A}$. For any $(x_1, \ldots, x_{m^4})$ the number of vectors in $\{(W^A(x_1), \ldots, W^A(x_{m^4})) \;;\; A \in \mathcal{A}\}$ is less than the number of vectors in

$\{(Z^A(x_1),\ldots,Z^A(x_{m^4}))\ ;\ A\in\mathcal{A}\}\le S(\mathcal{A},2m^4)$, where there the last bound follows from Proposition 4.1.8. Hence we can apply Devroye's result with $n=m^2$ and $S(\hat{\mathcal{A}},n^2)\le S(\mathcal{A},2m^4)$. □

The idea is to relax $\mathbb{P}_{\mathcal{A}}$ to $\mathbb{P}_{\mathcal{A},d}=\{(p_0,p_1,p_2)\ ;\ p_0+p_1+p_2=1, p_2\le\beta+d\}$ and because $\Pr\{\sup_{A\in\mathcal{A}}\hat{p}_2\le\beta+d\}\ge 1-4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}$ this occurs with high probability. In the remaining part we have no control over p_2 and we apply the existing bounding technique for general $\{0,1,2\}$-valued random variable.

As in Devroye's proof by using Hoeffding's inequality but now for $\{0,1,2\}$-valued random variables in Step 3 of Lemma 4.1.3, we can write

$$
\begin{aligned}
\Pr\left\{\left|\frac{1}{2m}\sum_{i=1}^m z_i-\widehat{P}_{(T,V)}(A)\right|>\underbrace{(1-\alpha)\epsilon\frac{m'}{m^2}}_{=\epsilon'}\right\}
&=\Pr\left\{\left|\frac{1}{m}\sum_{i=1}^m z_i-2\widehat{P}_{(T,V)}(A)\right|>2\epsilon'\right\}\\
&\le 2e^{-2m(2\epsilon')^2/2^2} && \text{by Hoeffding}\\
&=2e^{-2m\epsilon'^2}\\
&\le 2e^{-2m\epsilon^2+4\alpha m\epsilon^2+4\epsilon^2}\\
&=2e^{-2m\epsilon^2+4\epsilon+4\epsilon^2} && \text{when } \alpha=\frac{1}{m\epsilon}.
\end{aligned}
$$

By using the two parts obtained above, we can write

$$
\begin{aligned}
E\left[e^{-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*_{\bar{p}}(x)}\right]\le{}&\left(1-4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}\right)e^{-m\inf_{p\in\mathbb{P}_{\mathcal{A},d}}\inf_{x\in F_{\epsilon'}}\Lambda^*_{\bar{p}}(x)}\\
&+4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}e^{-2m\epsilon^2+4\epsilon+4\epsilon^2}.
\end{aligned}
$$

This together with Lemma 4.1.3 gives the statement of the theorem. □

t:sicont

Theorem 4.1.15. *For fixed $\epsilon>0$, given $\lambda>0$ there exists a ρ^0 such that*

$$\min\{\Lambda^*_{\bar{p}}(p_1+2p_2+2\epsilon),\Lambda^*_{\bar{p}}((p_1+2p_2-2\epsilon)\}\ge(8-\lambda)\epsilon^2$$

whenever $0<p_2\le\rho^0$, $p_1\ge 0$ and $p_1+p_2\le 1$.

Corollary 4.1.16. *Using Theorem 4.1.13 if $m > M$ such that $\epsilon' = (m\epsilon - 1)(m-1)/m^2 > \epsilon_M$ and $\beta + d < \rho^0$ then*

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))-P(A)\right|>\epsilon\right\}$$
$$\leq 4S(\mathcal{A},2m^2)\left[\left(1-C(d,m^2)\right)e^{-(8-\lambda)m\epsilon_M^2}+C(d,m^2)e^{-2m\epsilon^2+4\epsilon+4\epsilon^2}\right],$$

where the "correction term" $C(d,m^2) = 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}$.

Proof. First for $p_1 > 0$, $p_2 > 0$, $\min\{\Lambda^*_{p_1,p_2}(p_1+2p_2+2\epsilon), \Lambda^*_{p_1,p_2}(p_1+2p_2-2\epsilon\}$ is a continuous function on p_1 and p_2. We begin by studying the rate function as (p_1,p_2) approaches the corner $(0,0)$ or $(1,0)$.

First let $(p_1,p_2)\to(0,0)$. Then

$$\Lambda^*_{p_1,p_2}(p_1+2p_2+2\epsilon)$$
$$=\inf_{q_0,q_1,q_2}\{q_0\ln\frac{q_0}{p_0}+q_1\ln\frac{q_1}{p_1}+q_2\ln\frac{q_2}{p_2}\ ;\ q_1+2q_2=p_1+2p_2+2\epsilon,\sum q_i=1\}$$

and $q_1+2q_2\to 2\epsilon$ as $(p_1,p_2)\to(0,0)$. There exists $r < r_0 = \frac{\epsilon}{8}e^{-128/\epsilon}$ such that $p_1^2+p_2^2 < (2r)^2$ implies that $q_1 > \epsilon/2$ or $q_2 > \epsilon/4$. Then by using $x\ln(x/p) > x\ln x > -1/e$ for $p\in(0,1)$ and the fact that $q\ln(q/p) > 32$ when $q > \epsilon/4$ and $p < r_0$, we have

$$q_0\ln\frac{q_0}{p_0}+q_1\ln\frac{q_1}{p_1}+q_2\ln\frac{q_2}{p_2} > -\frac{2}{e}+32.$$

Hence $\Lambda^*_{p_1,p_2}(p_1+2p_2+2\epsilon) > 16$. Observe that as $(p_1,p_2)\to(0,0)$ we need not study $\Lambda^*_{p_1,p_2}(p_1+2p_2-2\epsilon)$ as by definition $\Lambda^*_{p_1,p_2}(x)=\infty$ if $x\notin[0,2]$.

Similarly, as $(p_1,p_2)\to(1,0)$, $q_1+2q_2 = p_1+2p_2+2\epsilon\to 1+2\epsilon$. Because $q_1\leq 1$, there exists $r < r_0$ such that $(1-p_1)^2+p_2^2 < (2r)^2$ implies that $q_2 > \epsilon/2$. Then again $q_0\ln\frac{q_0}{p_0}+q_1\ln\frac{q_1}{p_1}+q_2\ln\frac{q_2}{p_2} > -\frac{2}{e}+32$ and hence $\Lambda^*_{p_1,p_2}(p_1+2p_2+2\epsilon) > 16$.

Finally, when $q_1+2q_2 = p_1+p_2-2\epsilon\to 1-2\epsilon$ there exists $r < r_0$ such that $(1-p_1)^2+p_2^2 < (2r)^2$ implies that $q_1 < 1-\epsilon$. Hence either $q_0 > \epsilon/2$ or $q_2 > \epsilon/2$ and by repeating the above argument we conclude that $\Lambda^*_{p_1,p_2}(p_1+2p_2-2\epsilon) > 16$.

Now we take neighborhoods $V_{(0,0)} = \{(p_1,p_2)\ ;\ p_1^2+p_2^2 < (2r)^2, p_1+p_2 \le 1, p_1 > 0, p_2 > 0\}$ and $V_{(1,0)} = \{(p_1,p_2)\ ;\ (1-p_1)^2+p_2^2 < (2r)^2, p_1+p_2 \le 1, p_1 > 0, p_2 > 0\}$, where r is such that $(p_1,p_2) \in V_{(0,0)}$ implies that $\min\{\Lambda^*_{p_1,p_2}(p_1+2p_2+2\epsilon), \Lambda^*_{p_1,p_2}(p_1+2p_2-2\epsilon\} > 16 \ge 8\epsilon^2$, and $V_{(1,0)}$ similarly.

Fix p_1 such that $r \le p_1 \le 1-r$ and assume that $p_0 > 0$ and $p_2 > 0$. For $x \le 2$, $\Lambda^*_{p_1,p_2}(x) = F(x,\hat{q}(x))$, where

e:ent-f
$$F(x,q) = (1-x+q)\ln\frac{1-x+q}{1-p_1-p_2} + (x-2q)\ln\frac{x-2q}{p_1} + q\ln\frac{q}{p_2} \tag{4.4.4}$$

and $\hat{q}(x)$ is the unique solution in $(\max\{0,x-1\},x/2)$ of $\partial F/\partial q(x,\hat{q}) = 0$. $\partial F/\partial q(x,\hat{q}) = 0$ implies that

$$\frac{\hat{q}(x)(1+\hat{q}(x)-x)}{(x-2\hat{q}(x))^2} = \frac{p_0p_2}{p_1^2}.$$

Fix $p_1 > 0$ and as $p_2 \to 0$ either $\hat{q}(x) \to 0$ or $1+\hat{q}(x)-x \to 0$. The solution minimizing the relative entropy is $\hat{q}(x) \to 0$.

We claim that $\hat{q}(x) \to 0$ uniformly in p_1 provided that $p_1 \in [r,1-r]$. This can be seen by actually solving for $\hat{q}(x)$ by using the quadratic formula. Thus

$$\hat{q}(x) = \frac{-b \pm \sqrt{b^2-4ac}}{2a},$$

where $p = p_0p_2/p_1^2$, $a = 1-4p$, $b = 1-x+4xp$ and $c = -px^2$ and the correct sign is taken to give the solution in $(\max\{0,x-1\},x/2)$. Now $2a \ge r^2$ if $p_2 \le r^2/8$. For $p_1 \in [r,1-r]$, $p \le p_2/r^2$ and as $p_2 \to 0$, $p \to 0$ uniformly in p_1 and hence also $|\hat{q}(x)| \to 0$ as $p \to 0$. Hence $\hat{q}(x) \to 0$ uniformly in p_1 as $p_2 \to 0$, provided that $p_1 \in [r,1-r]$.

Let $x = p_1+2p_2+2\epsilon$. Then $\hat{q}(x) \to 0$, $1+\hat{q}(x)-x \to 1-(p_1+2\epsilon)$ and $x-\hat{q}(x) \to p_1+2\epsilon$ as $p_2 \to 0$. Thus

$$\begin{aligned}\Lambda^*_{p_1,p_2}(x) &= (1-x+\hat{q}(x))\ln\frac{1-x+\hat{q}(x)}{1-p_1-p_2} + (x-2\hat{q}(x))\ln\frac{x-2\hat{q}(x)}{p_1} + \hat{q}(x)\ln\frac{\hat{q}(x)}{p_2}\\ &\to (1-(p_1+2\epsilon))\ln\frac{1-(p_1+2\epsilon)}{p_0} + (p_1+2\epsilon)\ln\frac{p_1+2\epsilon}{p_1} = \Lambda^*_{p_2=0}(p_1+2\epsilon)\end{aligned}$$

uniformly in p_1 and $\Lambda^*_{p_1,p_2}(p_1+2p_2-2\epsilon)$ similarly.

We have previously shown that $\min\{\Lambda^*_{p_2=0}(p_1+2\epsilon), \Lambda^*_{p_2=0}(p_1-2\epsilon)\} \geq 8\epsilon^2$ and thus by taking p_2 small enough

$$\min\{\Lambda^*_{p_1,p_2}(p_1+2\epsilon), \Lambda^*_{p_1,p_2}(p_1-2\epsilon)\} \geq (8-\lambda)\epsilon^2.$$

If $p_1 = 0$,

$$\begin{aligned}\Lambda^*_{p_1=0}(x) &= \min\Big\{(1-q)\ln\frac{1-q}{1-p_2} + q\ln\frac{q}{p_2}\ ;\ 2q = x\Big\}\\ &= (1-\frac{x}{2})\ln\frac{1-\frac{x}{2}}{1-p_2} + \frac{x}{2}\ln\frac{\frac{x}{2}}{p_2}\end{aligned}$$

and

$$\Lambda^*_{p_1=0}(2p_2+2\epsilon) = (1-(p_2+\epsilon))\ln\frac{1-(p_2+\epsilon)}{1-p_2} + (p_2+\epsilon)\ln\frac{p_2+\epsilon}{p_2} \to \infty$$

as $p_2 \to 0$ establishing the result for $p_1 = 0$. □

Second order bound Our aim is to bound $\inf_{x\in F_{\epsilon'}} \Lambda^*_{\bar{p}}(x) \geq c\epsilon'^2$, where $c > 2$. For this we write

$$\Lambda^*_{\bar{p}}(p_1+2p_2+2\epsilon') = \Lambda^*_{\bar{p}}(p_1+2p_2) + \Lambda^{*\prime}_{\bar{p}}(p_1+2p_2)2\epsilon' + 1/2\Lambda^{*\prime\prime}_{\bar{p}}(\theta)(2\epsilon')^2,$$

where $\theta \in [p_1+2p_2, p_1+2p_2+2\epsilon']$. Again the properties of the rate function imply that $\Lambda^*_{\bar{p}}(p_1+2p_2) = \Lambda^{*\prime}_{\bar{p}}(p_1+2p_2) = 0$ and we aim at finding a lower bound for $\Lambda^{*\prime\prime}_{\bar{p}}(\theta)$.

For $x \leq 2$, $\Lambda^*_{\bar{p}}(x) = F(x,\hat{q}(x))$, where $F(x,q)$ is given by (4.4.4) and $\hat{q}(x)$ is the unique solution in $(\max(0,x-1), x/2)$ of $\partial F/\partial q(x,\hat{q}) = 0$. From this $\hat{q}'(x) = -F_{xq}(x,q)/F_{qq}(x,q)$ and $\Lambda^{*\prime}_{\bar{p}}(x) = F_x(x,\hat{q}) + F_q(x,\hat{q})\hat{q}' = F_x(x,\hat{q})$ and

$$\begin{aligned}\Lambda^{*\prime\prime}_{\bar{p}}(x) &= \frac{1}{2\hat{q}(x) + x(1-x)}\\ \hat{g}(x) &= \frac{x}{2} + \frac{1}{2\gamma}\left(1-\sqrt{1+\gamma(2x-x^2)}\right),\end{aligned}$$

if $\gamma = 4p_0p_2/p_1^2 - 1 \neq 0$ and $\hat{q}(x) = x^2/4$ if $\gamma = 0$.

By calculating $\Lambda_{\bar{p}}^{*'''}(x)$ we observe that $\Lambda_{\bar{p}}^{*'''}(1) = 0$ and for $\lambda = p_0 p_2 / p_1^2 > 1/16$ the minimum of $\Lambda_{\bar{p}}^{*''}(x)$ occurs at $x = 1$. Thus for $p_0 p_2 > p_1^2/16$ this gives

$$\min \Lambda_{\bar{p}}^{*''}(\theta) \geq \Lambda_{\bar{p}}^{*''}(1) = \frac{p_1^2 - 4p_2 + 4p_1p_2 + 4p_2^2}{-4p_2 + 4p_1p_2 + 4p_2^2 + \sqrt{p_1^2(-4p_1p_2 + 4(1-p_2)p_2)}}.$$

The key observation is that

$$\lim_{p_1 \to 0} \Lambda_{\bar{p}}^{*''}(1) = 1$$

yielding to bound $\Lambda_{\bar{p}}^{*}(p_1 + 2p_2 + 2\epsilon') \geq 2\epsilon'^2$. Thus by using the worst case analysis we are not able to indicate improvement from side information . This is in agreement with the analysis when $p_1 = 0$. This is the case in which $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0, 2\}$ (for example the side information mapping is $s(x) = x$), and we can not expect improvement from side information without further assumptions.

Next we analyze briefly the special case $p_1 = 0$. In this case

$$\begin{aligned} \Lambda_{p_1=0}^{*}(x) &= \min_q \{(1-q)\ln\frac{1-q}{1-p_2} + q\ln\frac{q}{p_2}\ ;\ 2q = x\} \\ &= (1 - x/2)\ln\frac{1-x/2}{1-p_2} + x/2\ln\frac{x/2}{p_2}. \end{aligned}$$

and $\Lambda_{\bar{p}}^{*''}(x) = \left(\frac{1}{2}\right)^2 \frac{1}{(1-x/2)x/2} \geq 1$. The minimum occurs when $x = 1$ (i.e., $q = 1/2$).

Note that for the 2nd order bounding technique the global minimum of the 2nd derivative is not needed; instead we are interested in

$$\inf\{\Lambda_{\bar{p}}^{*''}(\theta)\ ;\ \theta \in [p_1 + 2p_2 - 2\epsilon, p_1 + 2p_2 + 2\epsilon]\}.$$

Figure 4.1 illustrates this surface as a function of p_1 and p_2 for fixed $\epsilon = 0.1$. There we see that posing an upper bound for p_2 can improve the rate term. Also this illustrates the previous observation that the minimum of the second derivative occurs with small values of p_1.

For fixed ϵ and ρ, finding

$$\min_{\substack{p_1 \geq 0, p_2 \geq 0 \\ p_1 + p_2 \leq 1 \\ p_2 \leq \rho}} \left(\min_{\theta \in [p_1 + 2p_2 - 2\epsilon, p_1 + 2p_2 + 2\epsilon]} \Lambda_{\bar{p}}^{*''}(\theta) \right)$$

analytically is hard and we will not carry it further here.

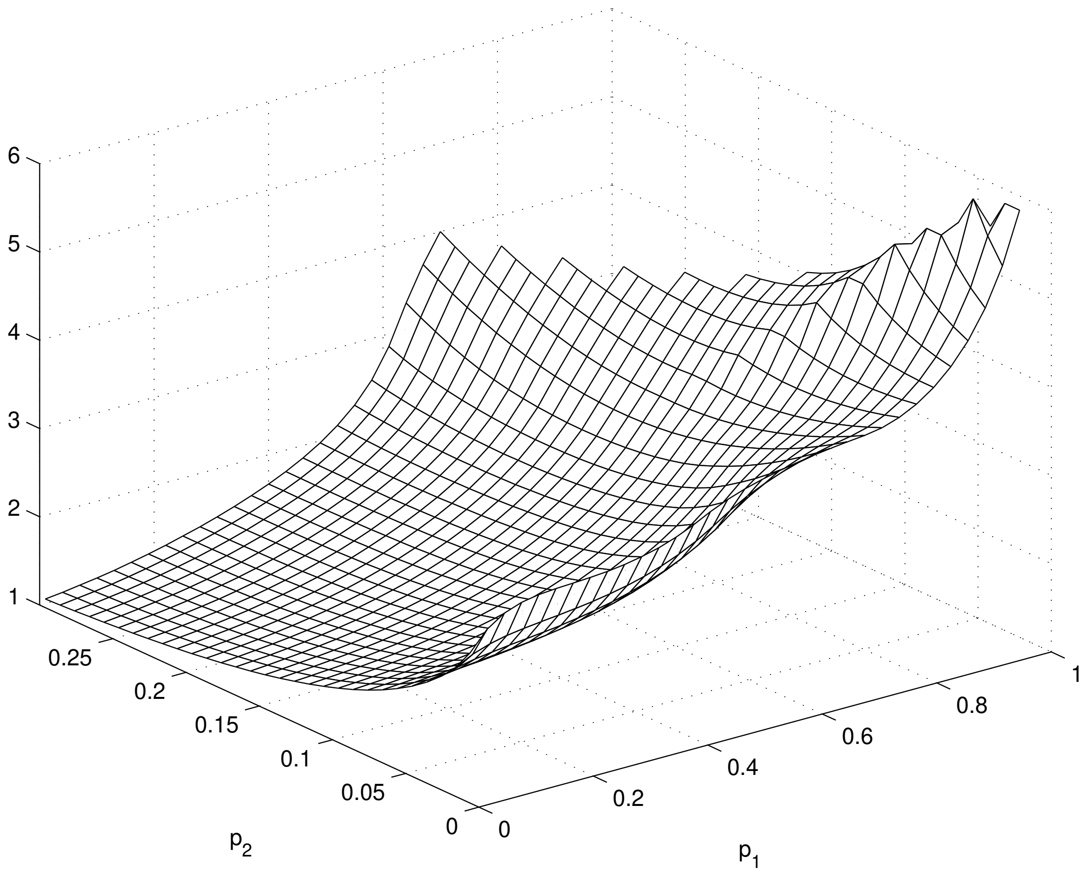

Figure 4.1: An illustration of the surface $\inf\{\Lambda_{\tilde{p}}^{*\prime\prime}(\theta)\ ;\ \theta \in [p_1+2p_2-2\epsilon, p_1+2p_2+2\epsilon]\}$ for $\epsilon = 0.1$ when $p_1 \in (0,1)$ and $p_2 \in (0, 0.3)$.

f:2der-surf

Third order bound We develop a 3rd order bound around the point $p_1 + 2p_2 = 1/2$ and we verify numerically that it is also a global bound. The advantage is that $\hat{q}(p_1+2p_2) = p_2$ and derivatives of $\Lambda^*_{\bar{p}}(x)$ at $x = p_1+2p_2$ are fairly simple expressions of p_1 and p_2. The idea is to find the minimum of the 2nd derivative and then bound the third order correction term. It turns out that the third order correction involves $\Lambda^{*'''}_{\bar{p}}(p_1 + 2p_2)$, which is still a fairly simple expression.

We want to calculate

$$\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon') = \Lambda^*_{\bar{p}}(p_1 + 2p_2) + \Lambda^{*'}_{\bar{p}}(p_1 + 2p_2)2\epsilon' \\ + (1/2)\Lambda^{*''}_{\bar{p}}(p_1 + 2p_2)(2\epsilon')^2 + (1/6)\Lambda^{*'''}_{\bar{p}}(\theta)(2\epsilon')^2,$$

where $\theta \in [p_1{+}2p_2, p_1{+}2p_2{+}2\epsilon']$ and $\Lambda^*_{\bar{p}}(p_1{+}2p_2{-}2\epsilon')$ similarly. Because $\hat{q}(p_1{+}2p_2) = p_2$ (this gives $q_0 = p_0$, $q_1 = p_1$ and $q_2 = p_2$ which has the relative entropy zero), we get easily

$$\Lambda^{*''}_{\bar{p}}(p_1 + 2p_2) = \frac{-1}{-p_1 + p_1^2 - 4p_2 + 4p_1p_2 + 4p_2^2} = \gamma(p_1, p_2).$$

If we fix p_2 then

$$\frac{\partial \gamma}{\partial p_1}(p_1, p_2) = \frac{-1 + 2p_1 + 4p_2}{(-p_1 + p_2^2 - 4p_2 + 4p_1p_2 + 4p_2^2)^2}$$

and we see that $\gamma(p_1, p_2)$ is minimized when $p_1 = 1/2 - 2p_2$.

Next assume that $p_2 \leq 1/4$ and $p_1 + 2p_2 = 1/2$. Now

$$\Lambda^{*''}_{\bar{p}}(1/2) = \Lambda^{*''}_{\bar{p}}(p_1 + 2p_2) = \frac{4}{1 + 8p_2}.$$

Also $\Lambda^{*'''}_{\bar{p}}(1/2) = \Lambda^{*'''}_{\bar{p}}(p_1{+}2p_2) = -192p_2/(1{+}8p_2)^3$. By graphing the 3rd derivative we verify that if $x \in [1/2, 1]$, $\Lambda^{*''}_{\bar{p}}(x) \geq -192p_2/(1 + 8p_2)^3$ and if $x \in [0, 1/2]$, $\Lambda^{*'''}_{\bar{p}}(x) \leq -192p_2/(1 + 8p_2)^3$. Thus for $\epsilon' \leq 1/4$, the third order correction term for $\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon')$ is greater than

$$\frac{1}{6}\frac{-192p_2}{(1 + 8p_2)^3}(2\epsilon')^3$$

and for $\Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon')$ the correction term is positive. Hence

$$\min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon'), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon')\} \geq \frac{1}{2}\frac{4}{1+8p_2}(2\epsilon')^2 - \frac{1}{6}\frac{192p_2}{(1+8p_2)^3}(2\epsilon')^3$$

for $p_1 + 2p_2 = 1/2$.

Define

$$R(p_2, \epsilon) = \frac{8}{1+8p_2} - \frac{8}{6}\frac{192p_2\epsilon}{(1+8p_2)^3}.$$

This is a decreasing function in p_2 ($\epsilon \leq 1/4$) and if we have a bound $p_2 \leq \rho$, then $R(p_2, \epsilon) \leq R(\rho, \epsilon)$.

By conducting an extensive numerical study we can verify that for $\epsilon \leq 1/4$ and $p_2 \leq \rho \in [0.032, 1/4]$

$$\min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon)\} > \frac{1}{2}\frac{4}{1+8p_2}(2\epsilon)^2 - \frac{1}{6}\frac{192p_2}{(1+8p_2)^3}(2\epsilon)^3.$$

First, the result for $\rho = 0.032$ was established by calculating

e:numcalc

$$\min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon)\} - R(\rho, \epsilon)\epsilon^2 \tag{4.4.5}$$

in a mesh where ϵ ranged from 0.0015 to 0.0015 + 0.25 in 100 steps and for each ϵ, p_2 ranged from 10^{-5} to ρ in 255 steps and p_1 ranged from 10^{-5} to $1 - p_2$ in 255 steps. This envolved over 6.5 million evaluations of (4.4.5) and we verified that $\min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon)\} > R(\rho, \epsilon)\epsilon^2$.

After that, the result was verified for ρ ranging from 0.032 to 0.25 in 30 steps. The range for ϵ remained the same but with 20 steps. When varying p_2 we used the fact that $R(\rho, \epsilon)$ is decreasing in ρ and it was only necessary to calculate p_2 from previously established ρ-value to the next ρ-value. This was done in 30 steps. For these p_2 values, p_1 ranged from 10^{-5} to $1 - p_2$ in 100 steps. Together this verification involved 2.4 million evaluations of (4.4.5).

The last verification was repeated again with increased step sizes; ρ had 40 steps, ϵ 30 steps, p_2 50 steps and p_1 120 steps. This envolved 7.2 million evaluations of (4.4.5) (that do not contain evaluations done previously, except for starting values) and the smallest difference was 3.13756×10^{-12}.

Another way to study the lower bound is to compare it directly to

$$\min_{\substack{0 \le p_2 \le \rho \\ 0 \le p_1 \le 1 \\ 0 \le p_1 + p_2 \le 1}} \min\{\Lambda^*_{\bar{p}}(p_1 + 2p_2 + 2\epsilon'), \Lambda^*_{\bar{p}}(p_1 + 2p_2 - 2\epsilon')\}$$

where the minimum has to be found numerically.

ϵ'	ρ	num. minimization	$R(\rho, \epsilon')(\epsilon')^2$
0.0901	0.15	0.02712	0.02688
0.0984	0.15	0.03213	0.03177
0.0901	0.1	0.03311	0.03286
0.0984	0.1	0.03922	0.03885

The above table shows a good agreement with two different methods and it indicates the the lower bound $R(\rho, \epsilon)\epsilon^2$ is quite tight.

Remark 4.1.17. By assuming that p_2 is bounded we inevitably pose a restriction on $P(A)$. However, as $p_1 + 2p_2 = P(A)$, assuming that $p_2 \le \rho$ implies that $P(A) \le 1/2 + \rho/2$. We can still compare the worst case analysis with and without side information, as the worst case without side information, $P(A) = 1/2$, is not excluded by our assumption on p_2.

Case 3: "Complete covering"

The case of "complete covering" in which $\bar{p}_A = (0, 1 - p, p)$ for each $A \in \mathcal{A}$ parallels the case $p_2 = 0$. Note, however, that the problem is not symmetric due to the relation $p_1 + 2p_2 = x$ in the relative entropy. Methods and ideas in the next two proofs correspond to ones in the case $p_2 = 0$ or p_2 small. We begin by showing an improvement in the rate function:

t:sqb-p0-0

Theorem 4.1.18. *Let $\mathcal{A}$ be a permissible class. If $\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i)) \in \{1,2\}$ for all i and $s : X \to X$ is measure preserving then*

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))-P(A)\right|\geq\epsilon\right\}$$
$$\leq 4e^{16\epsilon^2+16\epsilon}E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})\right]e^{-8m\epsilon^2}$$
$$\leq 4e^{16\epsilon^2+16\epsilon}S(\mathcal{A},2m^2)e^{-8m\epsilon^2}.$$

Proof. In this case the $\hat{p}=(\hat{p}_0,\hat{p}_1,\hat{p}_2)$ from the Step 3 in the proof of Lemma 4.1.3 satisfies $\hat{p}\in\mathbb{P}_{\mathcal{A}}$ and

$$E\left[e^{-m\inf_{x\in F_{\epsilon'}}\widetilde{\Lambda}^*_{\hat{p}}(x)}\right]\leq e^{-m\inf_{\bar{p}\in\mathbb{P}_{\mathcal{A}}}\inf_{x\in F_{\epsilon'}}\Lambda^*_{\bar{p}}(x)},$$

where

$$\Lambda^*_{\bar{p}}(x)=\Lambda^*_p(x)=\begin{cases}(2-x)\ln\frac{2-x}{p}+(x-1)\ln\frac{x-1}{1-p}, & \text{if } x\in[1,2],\\ \infty, & \text{otherwise.}\end{cases}$$

Because

$$\Lambda^{*\prime\prime}_p(x)=\frac{1}{-2+3x-x^2}\geq 4$$

we conclude that

$$\inf_{x\in F_{\epsilon'}}\Lambda^*_p(x)=\min\{\Lambda^*_p(p+2\epsilon'),\Lambda^*_p(p-2\epsilon')\}\geq 8{\epsilon'}^2.$$

As $e^{-8m{\epsilon'}^2}\leq e^{16\epsilon^2+16\epsilon}e^{-8m\epsilon^2}$, the result follows from Lemma 4.1.3. □

In the same way we can show improvement when p_0 is small:

Theorem 4.1.19. *For fixed $\epsilon>0$, given $\lambda>0$ there exists a ρ^0 such that*

$$\min\{\Lambda^*_{\bar{p}}(p_1+2p_2+2\epsilon),\Lambda^*_{\bar{p}}((p_1+2p_2-2\epsilon)\}\geq(8-\lambda)\epsilon^2$$

whenever $0<p_0\leq\rho^0$, $p_1\geq 0$ and $p_0+p_1\leq 1$.

Proof. This paralles the proof of Theorem 4.1.15. We begin by studying the corners $(p_0, p_1) = (0, 0)$ and $(p_0, p_1) = (0, 1)$.

First let $(p_0, p_1) \to (0, 0)$ and hence $p_2 \to 1$. Now $q_1 + 2q_2 = 2 - q_1 - 2q_0 = p_1 + 2p_2 - 2\epsilon \to 2 - 2\epsilon$. Hence $q_1 + 2q_0 \to 2\epsilon$ and $q_0 \not\to 0$ or $q_1 \not\to 0$. This implies that $\Lambda^*_{p_0,p_1}(p1 + 2p_2 - 2\epsilon) \to \infty$. The formal argument can be done as in the proof of Theorem 4.1.15 and it is omitted here.

Let $(p_0, p_1) \to (0, 1)$ and let $q_1 + 2q_2 = p_1 + 2p_2 + 2\epsilon \to 1 + 2\epsilon$. As $q_1 \le 1$, $q_2 \not\to 0$ and $\Lambda^*_{p_0,p_1}(p_1 + 2p_2 + 2\epsilon) \to \infty$. Similarly, if $q_1 + 2q_2 = p_1 + 2p_2 - 2\epsilon \to 1 - 2\epsilon$, then $1 - q_1 - 2q_0 \to -2\epsilon$. As $q_1 \le 1$ it is not possible that $q_0 \to 0$ and hence $\Lambda^*_{p_0,p_1}(p_1 + 2p_2 - 2\epsilon) \to \infty$.

As in the proof of Theorem 4.1.15 we take $\tilde r$ and the corresponding quarter balls $\widetilde V_{(0,0)}$ and $\widetilde V_{(0,1)}$ so that $(p_0, p_1) \in \widetilde V_{(0,0)} \cup \widetilde V_{(0,1)}$ implies that

$$\min\{\Lambda^*_{p_1,p_2}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{p_1,p_2}(p_1 + 2p_2 - 2\epsilon\} > 16 \ge 8\epsilon^2.$$

Fix p_1 such that $\tilde r \le p_1 \le 1 - \tilde r$ and assume that $p_0 > 0$ and $p_2 > 0$. For $0 \le x \le 2$ as

$$\begin{aligned}
&\Lambda^*_{\bar p}(x)\\
&= \Lambda^*_{p_0,p_1}(x) = \inf_{q_0,q_1,q_2}\Big\{q_0 \ln\frac{q_0}{p_0} + q_1 \ln\frac{q_1}{p_1} + q_2 \ln\frac{q_2}{p_2}\ ;\ q_1 + 2q_2 = x, q_0 + q_1 + q_2 = 1\Big\}\\
&= \inf_{\substack{\max(0,x-1)\le r\\ \le 1-x/2}}\Big\{r \ln\frac{r}{p_0} + (2 - 2r - x)\ln\frac{2 - 2r - x}{p_1} + (x - 1 + r)\ln\frac{x - 1 + r}{p_2}\Big\},
\end{aligned}$$

where $p_2 = 1 - p_0 - p_1$.

Assume $p_0 > 0$, $p_1 > 0$ and $p_2 > 0$. For $x \le 2$, $\Lambda^*_{p_0,p_1}(x) = G(x, \hat r(x))$, where

e:ent-g
$$G(x, r) = r \ln\frac{r}{p_0} + (2 - 2r - x)\ln\frac{2 - 2r - x}{p_1} + (x - 1 + r)\ln\frac{x - 1 + r}{p_2} \qquad (4.4.6)$$

and $\hat r(x)$ is the unique solution in $(\max(0, x - 1), 1 - x/2)$ of $\partial G/\partial r(x, \hat r) = 0$. $\partial G/\partial r(x, \hat r) = 0$ implies that

$$\frac{\hat r(x)(x - 1 + \hat r(x))}{(2 - 2\hat r(x) - x)^2} = \frac{p_0 p_2}{p_1^2}.$$

Fix $p_1 > 0$ and as $p_0 \to 0$ either $\hat{r}(x) \to 0$ or $1 + \hat{r}(x) - x \to 0$. Again the solution minimizing the relative entropy is $\hat{r}(x) \to 0$. By an argument similar to one in Theorem 4.1.15 we can show that $\hat{r}(x) \to 0$ uniformly in p_1.

Let $x = p_1 + 2p_2 + 2\epsilon = 2 - 2p_0 - p_1 + 2\epsilon$. Then $\hat{r}(x) \to 0$, $x - 1 + \hat{r}(x) \to 1 - p_1 + 2\epsilon$ and $2 - 2\hat{r}(x) - x \to p_1 - 2\epsilon$ as $p_0 \to 0$. Now $\Lambda^*_{p_0,p_1}(x)$ equals

$$\hat{r}(x) \ln \frac{\hat{r}(x)}{p_0} + (2 - 2\hat{r}(x) - x) \ln \frac{2 - 2\hat{r}(x) - x}{p_1} + (x - 1 + \hat{r}(x)) \ln \frac{x - 1 + \hat{r}(x)}{p_2},$$

which approaches

$$(p1 - 2\epsilon) \ln \frac{p_1 - 2\epsilon)}{p_1} + (1 - p_1 + 2\epsilon) \ln \frac{1 - p_1 + 2\epsilon}{1 - p_1} = \Lambda^*_{p_0=0}(2 - p_1 + 2\epsilon)$$

uniformly in p_1 provided that $p_1 \in [\tilde{r}, 1 - \tilde{r}]$ and $\Lambda^*_{p_0,p_1}(p_1 + 2p_2 - 2\epsilon)$ similarly.

We have previously shown that $\min\{\Lambda^*_{p_0=0}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{p_0=0}(p_1 + 2p_2 - 2\epsilon)\} \geq 8\epsilon^2$ and thus by taking p_0 small enough

$$\min\{\Lambda^*_{p_0,p_1}(p_1 + 2p_2 + 2\epsilon), \Lambda^*_{p_0,p_1}(p_1 + 2p_2 - 2\epsilon)\} \geq (8 - \lambda)\epsilon^2.$$

If $p_1 = 0$, take $x = 2p_2 - 2\epsilon$. As

$$\Lambda^*_{p_1=0}(x) = (1 - x/2) \ln \frac{1 - x/2}{p_0} + x/2 \ln \frac{x/2}{1 - p_0},$$

we have

$$\Lambda^*_{p_1=0}(2 - 2p_0 - 2\epsilon) = (p_0 + \epsilon) \ln \frac{p_0 + \epsilon}{p_0} + (1 - p_0 - \epsilon) \ln \frac{1 - p_0 - \epsilon}{1 - p_0} \to \infty$$

as $p_0 \to 0$, establishing the result for $p_1 = 0$. $\square$

Further, one can expect a third order analysis comparable to the case p_2 small to work in terms of p_0 and p_1. Now $\hat{q}(p_1 + 2p_2) = p_2$ translates to $\hat{r}(2 - p_1 - 2p_0) = p_0$. In terms of p_0 and p_1, the minimum of the second derivative occurs at $p_1 + 2p_0 = 1/2$, but the third order analysis is around the point $x = 3/2$ (instead of $x = 1/2$). One expects that $p_1 + 2p_2 - 2\epsilon$ has the negative correction term.

4.1.4 Numerical Comparison of Side Information Improvement

We briefly illustrate how our results compare to the bound by Devroye:

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|\geq\epsilon\right\}\leq 4e^{4\epsilon+4\epsilon^2}S(\mathcal{A},m^2)e^{-2m\epsilon^2}.$$

First if there is no overlap, we improve the exponential term from $e^{-2m\epsilon^2}$ to $e^{-8m\epsilon^2}$ (ignoring small e^{ϵ} type of terms) with a small increase in a term that grows polynomially in m.

Next we give examples of improvement when p_2 is small using Theorem 4.1.13 together with the numerically verified lower bound (4.4.5) with $\epsilon' = (m\epsilon-1)(m-1)/m^2$ and we bound $S(\mathcal{A},m)$ by Sauer's lemma, i.e., $S(\mathcal{A},m)\leq(em/\mathrm{VC}(\mathcal{A}))^{\mathrm{VC}(\mathcal{A})}$. We fix $\mathrm{VC}(\mathcal{A})$, ϵ, $p_2\leq\beta$, d (so $\rho=\beta+d$) and δ, in order to ask how many samples are needed so that

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|\geq\epsilon\right\}\leq\delta$$

and

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))-P(A)\right|\geq\epsilon\right\}\leq\delta.$$

If $\mathrm{VC}(A)=2$, $\epsilon=0.1$, $\delta=0.2$, $\beta=0.1$ and $d=0.05$, $m\geq 1689$ samples are needed without side information and $m\geq 1019$ samples with side information. Similarly, if $\delta=0.02$, the corresponding sample complexities are 1819 without side information and 1099 with side information.

By taking β smaller the exponential term improves. With $\mathrm{VC}(A)=2$, $\epsilon=0.1$, $\beta=0.05$ and $d=0.05$ we get sample complexities 815 for $\delta=0.2$ and 880 for $\delta=0.02$. Compared to sample complexities without side information roughly only half as many samples are needed.

We can still detect improvement when β increases to 0.2. With other parameters as before, we get the sample complexity 1422 for $\delta=0.2$ and 1529 for $\delta=0.02$.

4.1.5 Improved Exponential Bounds without Side Information

We illustrate very briefly how the bounding technique can be applied in problems without side information, but having some information on possible values for $P(A)$. Recall, that assumption on p_2 inevitably translates to an assumption of $P(A)$. However, that assuption does not exclude the worst case of learning without side information. Next our aim is to sketch how a stronger assumption can lead to improved exponential rates without side information, but those rates are, in general, worse than the ones with side information (if it is available).

Without side information we can repeat the same analysis but now the exponential term is given by $e^{-m \inf_{x \in F_{\epsilon'}} \Lambda_p^*(x)}$, where $p = P(A)$, $F_{\epsilon'} = (-\infty, p - \epsilon'] \cup [p + \epsilon', \infty)$ and

$$\Lambda_p^*(x) = \begin{cases} (1-x) \ln \frac{1-x}{1-p} + x \ln \frac{x}{p}, & \text{if } x \in [0,1], \\ \infty, & \text{otherwise.} \end{cases}$$

As before $\inf_{x \in F_{\epsilon'}} \Lambda_p^*(x) = \min\{\Lambda_p^*(p-\epsilon'), \Lambda_p^*(p+\epsilon')\}$. Write $\Lambda_p^*(p+\epsilon') = \Lambda_p^*(p) + \Lambda_p^{*\prime}(p)\epsilon' + 1/2 \Lambda_p^{*\prime\prime}(\theta)\epsilon'^2 = 1/2 \Lambda_p^{*\prime\prime}(\theta)\epsilon'^2$, where $\theta \in [p, p+\epsilon']$. Now

$$\Lambda_p^{*\prime\prime}(x) = \frac{1}{x(1-x)} \text{ for } x \in (0,1).$$

Assume that we know $P(A) \leq \lambda$. As with the case p_2 small we need to relax our class of probability vectors from $\{(p, 1-p) \; ; \; p \leq \lambda\}$ to $\{(p, 1-p) \; ; \; p \leq \lambda + d\}$ for some $d > 0$. Assuming for simplicity that $\lambda + d + \epsilon' < 1/2$ we get a bound for the second derivative

$$\frac{1}{(1-\lambda-d-\epsilon')(\lambda+d+\epsilon')} = g(\lambda+d, \epsilon') > 4.$$

and the learning bound becomes

$$\Pr\left\{\sup_{A\in\mathcal{A};P(A)\le\lambda}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|\ge\epsilon\right\}$$
$$\le 4S(\mathcal{A},m^2)\left(\left(1-4e^{4d^2+4d}S(\mathcal{A},m^4)e^{-2d^2m^2}\right)e^{-1/2g(\lambda+d,\epsilon')\epsilon'^2}\right.$$
$$\left.+\,4e^{4d^2+4d}S(\mathcal{A},m^4)e^{-2d^2m^2}e^{-2m\epsilon'^2}\right)$$

where $\epsilon' = (m\epsilon-1)(m-1)/m^2$. Here the correction term is derived as in Proposition 4.1.14, the only difference is that no new concept class needs to be introduced. And again, we emphasize that the sample size in the correction term is m^2 making it quickly negligiable.

We want to point out that exponential terms obtained in this way are, in general, worse than the worst case terms with side information unless $\lambda + d + \epsilon'$ is extremely small. Also the above knowledge can be utilized with side information too (the easiest case is $p_2 = 0$ together with the knowledge $P(A) \le \lambda$) leading to improved bounds.

4.2 Bounds of the Form $C_1 e^{-c_2 m\epsilon}$

s:epsbounds

In this section we discuss how side information can improve bounds with exponential term $e^{-cm\epsilon}$. The treatment is similar to the case with ϵ^2 bounds in introducing the ghost sample and symmetrization but the exponential term comes from calculating a bound for the number of permutations satisfying a certain condition.

Shawe-Taylor et. al. [33] have shown that

$$\Pr\left\{\sup_{A\in\mathcal{A};\sum\mathbf{1}_A(X_i)=0}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i)-P(A)\right|>\epsilon\right\}\le 2\left(\frac{em}{2d}\right)^d\epsilon^d e^{2\sqrt{2d}}e^{-m\epsilon},$$

where $m \ge 4d/\epsilon$ and $d = \mathrm{VC}(\mathcal{A})$. Their proof follows the standard technique but it takes the parameters α and m' (that appeared also in the proof of the $e^{-cm\epsilon^2}$

bound) to depend on the VC-dimension of $\mathcal{A}$. By a simple choice $m' = m^2 - m$ and $\alpha = 1/m$ one arrives at a bound [18, p. 210]

$$\Pr\left\{ \sup_{A\in\mathcal{A};\sum \mathbf{1}_A(X_i)=0} \left| \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}_A(X_i) - P(A) \right| > \epsilon \right\} \le 2S(\mathcal{A}, m^2)e^{-m\epsilon}.$$

Our approach is to make the choice of m' and α as in the previous section and utilize a result proven there. This simplifies the process without changing the exponential term, although the constant term may not be the best possible (we get some terms of the type e^{ϵ}). We begin by proving a lemma setting up the problem and then separately we discuss when side information can improve the exponential term.

We review the notation briefly. Let $X_i \in X$ and let $s : X \to X$ be measure preserving. For a given $A \in \mathcal{A}$ we form $Z_i^A = 1_A(X_i) + 1_A(s(X_i)) = j$ with probability p_j for $j = 0, 1, 2$. We denote $T = X_1, \ldots, X_m$ and $V = X_{m+1}, \ldots, X_{m+m'}$, where $m' = m^2 - m$. Now $Z_1^A, \ldots, Z_{m^2}^A$ are i.i.d. random variables and let

$$\widehat{P}_T(A) = \frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = \frac{1}{2m}\sum_{i=1}^{m} Z_i^A \text{ and}$$

$$\widehat{P}_V(A) = \frac{1}{2m'}\sum_{i=1}^{m'} \mathbf{1}_A(X_{m+i}) + \mathbf{1}_A(s(X_{m+i})) = \frac{1}{2m'}\sum_{i=1}^{m'} Z_{m+i}^A.$$

l:epsgenbnd

Lemma 4.2.1. *Assume that the class $\mathcal{A}$ is permissible. Then*

$$\Pr\{ \sup_{\substack{A\in\mathcal{A}\\ \widehat{P}_T(A)=0}} |\widehat{P}_T(A) - P(A)| \ge \epsilon\} \le 2\Pr\{ \sup_{\substack{A\in\mathcal{A}\\ \widehat{P}_T(A)=0}} |\widehat{P}_T(A) - \widehat{P}_V(A)| \ge (1-\alpha)\epsilon\}$$

$$\le 2E\left[\widetilde{N}_{\mathcal{A}}(X_1, \ldots, X_{m^2}) e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D \right] \le 2\widetilde{S}(\mathcal{A}, m^2) E\left[e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D \right],$$

where

$$
\begin{aligned}
l^A &= \text{ number of 1's in } z_1^A, \ldots, z_{m^2}^A,\\
k^A &= \text{ number of 2's in } z_1^A, \ldots, z_{m^2}^A,\\
D &= \{m^2 - m \geq \ell^A + k^A > 2(m^2-m)(1-\alpha)\epsilon - k^A\},\\
\alpha &= 1/(m\epsilon),\\
\widetilde{N}_{\mathcal{A}}(X_1, \ldots, X_{m^2}) &= |\{(Z_1^A, \ldots, Z_{m^2}^A) \ ; \ A \in \mathcal{A}\}| \text{ and}\\
\widetilde{S}(\mathcal{A}, m) &= \max_{(x_1,\ldots,x_m)} |\{(\mathbf{1}_A(x_1) + \mathbf{1}_A(s(x_1)), \ldots, \mathbf{1}_A(x_m) + \mathbf{1}_A(s(x_m))) \ ; \\
&\qquad A \in \mathcal{A}\}|.
\end{aligned}
$$

Proof. We start the 4 step process utilizing Step 1 in the proof of Lemma 4.1.3.

Step 1: A ghost sample of size $m' = m^2 - m$

We only redefine the events

$$
\begin{aligned}
\{\sigma > \epsilon\} &= \left\{ \sup_{A; \widehat{P}_T(A)=0} |\widehat{P}_T(A) - P(A)| > \epsilon \right\}\\
\{\rho > (1-\alpha)\epsilon\} &= \left\{ \sup_{A; \widehat{P}_T(A)=0} |\widehat{P}_T(A) - \widehat{P}_V(A)| > (1-\alpha)\epsilon \right\}
\end{aligned}
$$

and then like in the proof of Step 1 in Lemma 4.1.3 we show that for all $0 < \alpha < 1$ and $\epsilon > 0$,

$$\Pr\{\rho > (1-\alpha)\epsilon\} \geq \left(1 - \frac{1}{4\alpha^2\epsilon^2 m'}\right) \Pr\{\sigma > \epsilon\}.$$

Step 2: Symmetrization by permuting

The distribution of $\sup_{A; \widehat{P}_T(A)=0} |\widehat{P}_T(A) - \widehat{P}_V(A)|$ is the same as the distribution of

$$\beta(\pi) := \sup_{A; \sum_{i=1}^m (\pi Z_i)^A = 0} \left| \frac{1}{2m} \sum_{i=1}^{m} (\pi Z_i)^A - \frac{1}{2m'} \sum_{i=1}^{m'} (\pi Z_{m+i})^A \right|,$$

where π is a permutation of $m + m' = m^2$ random variables $X_1, \ldots, X_m, X_{m+1}, \ldots, X_{m+m'}$ and by $(\pi Z_i)^A$ we denote $\mathbf{1}_A(\pi(X_i)) + \mathbf{1}_A(s(\pi(X_i)))$.

There are $m^2!$ possible permutations $\pi_1, \ldots, \pi_{m^2!}$. Thus

$$\Pr\Big\{ \sup_{A;\sum_{i=1}^m (\pi Z_i)^A = 0} |\widehat{P}_T(A) - \widehat{P}_V(A)| > (1-\alpha)\epsilon \Big\} = E\left[\frac{1}{m^2!} \sum_{j=1}^{m^2!} \mathbf{1}_{\{\beta(\pi_j) > (1-\alpha)\epsilon\}} \right].$$

Step 3: Conditioning

Fix $x_1, \ldots, x_m, x_{m+1}, \ldots, x_{m+m'}$ and let $\bar{\mathcal{A}} \subset \mathcal{A}$ be a collection of sets such that any two sets in $\bar{\mathcal{A}}$ give rise to different vectors $(z_1^A, \ldots, z_m^A, z_{m+1}^A, \ldots, z_{m+m'}^A)$. Now we can take the supremum over $\bar{\mathcal{A}}$ instead of over $\mathcal{A}$. To ease the notation for a moment let $P_{m,j} = 1/(2m) \sum_{i=1}^m (\pi_j z_i)^A$ and $P_{m',j} = 1/(2m') \sum_{i=1}^{m'} (\pi_j z_{m+i})^A$. Then we can bound

$$\begin{aligned}
&\frac{1}{m^2!} \sum_{j=1}^{m^2!} \sup_{A;P(m,j)=0} \mathbf{1}_{\{|P(m,j) - P(m',j)| > (1-\alpha)\epsilon\}} \\
&\quad = \frac{1}{m^2!} \sum_{j=1}^{m^2!} \sup_{A\in\bar{\mathcal{A}};P(m,j)=0} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}} \\
&\quad \le \frac{1}{m^2!} \sum_{j=1}^{m^2!} \sum_{A\in\bar{\mathcal{A}};P(m,j)=0} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}} \\
&\quad = \frac{1}{m^2!} \sum_{j=1}^{m^2!} \sum_{A\in\bar{\mathcal{A}}} \mathbf{1}_{\{P(m,j)=0\}} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}} \\
&\quad = \sum_{A\in\bar{\mathcal{A}}} \frac{1}{m^2!} \sum_{j=1}^{m^2!} \mathbf{1}_{\{P(m,j)=0\}} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}}.
\end{aligned}$$

Step 4: Counting

Fix a set $A \in \bar{\mathcal{A}}$ and observe that

$$\frac{1}{m^2!} \sum_{j=1}^{m^2!} \mathbf{1}_{\{P(m,j)=0\}} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}}$$

is the fraction of permutations satisfying

e:a-perm-def

$$\mathbf{1}_{\{P(m,j)=0\}} \mathbf{1}_{\{P(m',j) > (1-\alpha)\epsilon\}} = 1. \tag{4.4.7}$$

Let l^A and k^A denote the number of 1's and 2's, respectively, in the sequence $z_1^A, \ldots, z_m^A, z_{m+1}^A, \ldots, z_{m+m'}^A$, where $m + m' = m^2$. We call a permutation admissible if it satisfies (4.4.7). There are no admissible permutations if $l^A + k^A \geq m'$ or $(l^A+2k^A)/(2m') \leq (1-\alpha)\epsilon$. Hence we require that $m' \geq l^A+k^A > 2m'(1-\alpha)\epsilon - k^A$.

The fraction of permutations satisfying the above condition is bounded by

$$\begin{aligned}
\frac{\binom{m'}{k^A}\binom{m'-k^A}{l^A}}{\binom{m+m'}{k^A}\binom{m+m'-k^A}{l^A}} &\leq \left(\frac{m'}{m+m'}\right)^{k^A}\left(\frac{m'-k^A}{m+m'-k^A}\right)^{l^A} \\
&= \left(1-\frac{m}{m+m'}\right)^{k^A}\left(1-\frac{m}{m+m'-k^A}\right)^{l^A} \\
&\leq \exp\left(-\frac{m}{m+m'}k^A\right)\exp\left(-\frac{m}{m+m'-k^A}l^A\right) \\
&= \exp\left(\frac{-k^A}{m}\right)\exp\left(\frac{-ml^A}{m^2-k^A}\right),
\end{aligned}$$

where in the last line we have substituted $m' = m^2 - m$.

Now we can collect all the steps together. Note that the cardinality of $\bar{\mathcal{A}}$ is given by $\widetilde{N}_{\mathcal{A}}(X_1, \ldots, X_{m^2})$ which is in turn bounded by $\widetilde{S}(\mathcal{A}, m^2)$, see Propostition 4.1.8. We get

e:adm-perm

$$\begin{aligned}
&\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \widehat{P}_T(A)=0}} |\widehat{P}_T(A) - P(A)| > \epsilon\Big\} \leq 2\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \widehat{P}_T(A)=0}} |\widehat{P}_T(A) - \widehat{P}_V(A)| > (1-\alpha)\epsilon\Big\} \\
&\leq 2E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2}) e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D\right] \leq 2\widetilde{S}(\mathcal{A}, m^2) E\left[e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \mathbf{1}_D\right], \quad (4.4.8)
\end{aligned}$$

where $D = \{m^2 - m \geq l^A + k^A > 2(m^2-m)(1-\alpha)\epsilon - k^A\}$. □

We begin by showing that side information can double the sample size in the exponential term improving the best exponent $e^{-m\epsilon}$ to $e^{-2m\epsilon}$. Recall that $\widetilde{S}(\mathcal{A}, m) \leq S(\mathcal{A}, 2m)$.

t:epsb1

Theorem 4.2.2. *Assume that the class $\mathcal{A}$ is permissible. If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in$*

$\{0,1\}$ for all i and $s : X \to X$ is measure preserving then

$$\Pr\Big\{ \sup_{\substack{A\in\mathcal{A}\\ \sum_{i=1}^m Z_i^A=0}} |\frac{1}{2m}\sum_{i=1}^m Z_i^A - P(A)| > \epsilon\Big\}$$
$$\leq 2e^{2(1+\epsilon)}e^{-2m\epsilon}E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})\right] \leq 2S(\mathcal{A},2m^2)e^{2(1+\epsilon)}e^{-2m\epsilon}.$$

Proof. If $p_2 = 0$ then $k^A = 0$ and to obtain admissible permutations in the counting step of Lemma 4.2.1, $l^A > 2m'(1-\alpha)\epsilon$. By substituting $\alpha = 1/(m\epsilon)$ and $m' = m^2-m$ to (4.4.8) and simplifying we obtain the result. □

Next we indicate improvement when p_2 is small.

t:epsb2

Theorem 4.2.3. *Assume that $\mathcal{A}$ is permissible and $s : X \to X$ is measure preserving. If $\mathbb{P}_{\mathcal{A}} = \{(p_0,p_1,p_2)\ ;\ p_2 \leq \gamma\}$ and $\gamma < \epsilon$ then*

$$\Pr\Big\{ \sup_{\substack{A\in\mathcal{A}\\ \sum_{i=1}^m Z_i^A=0}} |\frac{1}{2m}\sum_{i=1}^m Z_i^A - P(A)| > \epsilon\Big\}$$
$$\leq 2E\left[\widetilde{N}_{\mathcal{A}}(X_1,\ldots,X_{m^2})\right]\Big[\Big(1-4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}\Big)e^{-mf(m^2\rho)}$$
$$+\ 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}e^{1+\epsilon}e^{-m\epsilon}\Big]$$
$$\leq 2S(\mathcal{A},2m^2)\Big[\Big(1-4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}\Big)e^{-mf(m^2\rho)}$$
$$+\ 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}e^{1+\epsilon}e^{-m\epsilon}\Big],$$

where

$$f(k) = \frac{k}{m^2} + \frac{2((m\epsilon-1)(m-1)-k)}{m^2-k},$$

$\rho = \gamma + d$ and $m^2\rho \leq (m\epsilon-1)(m-1)$.

Proof. Let l^A and k^A denote the number of 1's and 2's in the sequence $z_1^A,\ldots,z_m^A,z_{m+1}^A,\ldots,z_{m+m'}^A$. By taking $m' = m^2-m$ and $\alpha = 1/(m\epsilon)$ the condition for admissible permutations in Step 4 of the proof of Lemma 4.2.1 translates to condition $l^A + 2k^A > 2(m\epsilon-1)(m-1)$.

Then the upper bound for the fraction of admissible permutations becomes

$$e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}} \le e^{\frac{-k^A}{m}} e^{\frac{-2m}{m^2-k^A}\left((m\epsilon-1)(m-1)-k^A\right)} = e^{-mf(k^A)},$$

where

$$f(k) = \frac{k}{m^2} + \frac{2((m\epsilon-1)(m-1)-k)}{m^2-k}$$

for $0 \le k \le (m\epsilon - 1)(m-1)$. $f(k) > 0$ and f is a decreasing function of k. Note that $f(0)$ gives the bound of the previous theorem.

If $p_2 \le \gamma$, then for $d > 0$ such that $\rho = \gamma + d < \epsilon$, Proposition 4.1.14 gives that $\Pr\{\sup_{A\in\mathcal{A}} k^A/m^2 - \gamma > d\} \le 4e^{4d^2+4d} S(\mathcal{A}, 2m^4) e^{-2d^2m^2}$.

Thus

$$E\left[e^{\frac{-k^A}{m}} e^{\frac{-ml^A}{m^2-k^A}}\right] \le \left(1 - 4e^{4d^2+4d} S(\mathcal{A}, 2m^4) e^{-2d^2m^2}\right) e^{-mf(m^2\rho)} \\ + 4e^{4d^2+4d} S(\mathcal{A}, 2m^4) e^{-2d^2m^2} e^{1+\epsilon} e^{-m\epsilon}.$$

This with Lemma 4.2.1 implies the result. □

Note again that the sample size in the correction term $4e^{4d^2+4d} S(\mathcal{A}, 2m^4) e^{-2d^2m^2}$ is m^2 and it decreases very rapidly making $e^{-mf(m^2\rho)}$ the dominating term.

We can not indicate exponential improvement when p_2 is of size ϵ or larger. Assume for a moment that $p_2 = \epsilon$. Then the expected value of k^A is $m^2\epsilon$. Observe that if $k^A = m^2\epsilon$ and $l^A = 0$, the exponential term $e^{-k^A/m} e^{-ml^A/(m^2-k^A)} = e^{-m\epsilon}$ and we are not able improve from the best known exponent. Without a stronger assumption on p_2 we can not indicate a better exponential term. A situation like above occurs in the interval problem when the length of the target interval $T = (a, b)$ is $b - a = 1/2 + \epsilon/2$ and the side information mapping is $s(x) = (x + 1/2) \mod 1$. The learning algorithm is the smallest interval containing positive samples and hence for each estimate $\widetilde{h}_m$ the p_2 corresponding to $\widetilde{h}_m \Delta T$ satisfies $p_2 \le \epsilon$. Recall that the exact calculations proved the learning rate $(1-\epsilon)^m$. If one wants to exclude this case from discussion a more complicated relation for (p_0, p_1, p_2) has to be posed and analyzed.

4.3 Relaxing the Measure Preservability Assumption

s:rel-meas

We illustrate how the measure preservability assumption on the side information mapping $s(x) : X \to X$ can be relaxed. It turns out that in this case the analysis for both kinds of exponential bounds can be repeated with a smaller value of ϵ.

Assume that there exists a probability measure Q on X such that the side information mapping $s : X \to X$ is measure preserving with respect to Q. Assume further that P is absolutely continuous with respect to Q so that the Radon-Nikodym derivative $dP/dQ = f$ satisfies, for some constants c_1, c_2, $0 < c_1 \leq f \leq c_2$ for all $x \in X$. Thus Q is also absolutely continuous with respect to P and $dP/dQ = 1/f$.

Then

$$P(s^{-1}(A)) = \int_{s^{-1}(A)} f dQ \leq c_2 Q(A) \leq \frac{c_2}{c_1} P(A)$$

and similarly

$$P(s^{-1}(A)) = \int_{s^{-1}(A)} f dQ \geq c_1 Q(A) \geq \frac{c_1}{c_2} P(A).$$

Bounds for Consistent Algorithms

Assume that there exists a constant $\rho > 0$ such that $P(A) > \epsilon$ implies $P(s^{-1}(A)) > \rho\epsilon$, and then

$$\Big\{ \sup_{\substack{A \in \mathcal{A} \\ \sum \mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) = 0}} \Big| \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(X_i) + \frac{1}{m} \sum_{i=1}^{m} \mathbf{1}_A(s(X_i)) - 2P(A) \Big| > 2\epsilon \Big\}$$

is contained in

$$\Big\{\sup_{\substack{A\in\mathcal{A}\\ \sum \mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))=0}} \Big|\frac{1}{m}\sum_{i=1}^m \mathbf{1}_A(X_i)+\frac{1}{m}\sum_{i=1}^m \mathbf{1}_A(s(X_i))-P(A)-P(s^{-1}(A))\Big| > \epsilon+\rho\epsilon\Big\}.$$

Hence we can study

$$\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \sum Z_i^A=0}} \Big|\frac{1}{m}\sum_{i=1}^m Z_i^A - \big(P(A)+P(s^{-1}(A))\big)\Big| > \epsilon(1+\rho)\Big\}$$

in place of

$$\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \sum Z_i^A=0}} \Big|\frac{1}{m}\sum_{i=1}^m Z_i^A - 2P(A)\Big| > 2\epsilon\Big\},$$

and the analysis of $C_1e^{-c_2m\epsilon}$ bounds can be repeated with $\frac{1+\rho}{2}\epsilon$ in place of ϵ. This means that the improvement from side information is smaller if $s(x)$ is not measure preserving. For example, Theorem 4.2.2 translates to:

Theorem 4.3.1. *Assume that the class $\mathcal{A}$ is permissible. If $\mathbf{1}_A(X_i)+\mathbf{1}_A(s(X_i))\in\{0,1\}$ for all i and $s: X\to X$ is such that for some $\rho>0$, $P(A)>\epsilon$ implies $P(s^{-1}(A))>\rho\epsilon$ for all $A\in\mathcal{A}$ then*

$$\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \sum_{i=1}^m Z_i^A=0}} \Big|\frac{1}{2m}\sum_{i=1}^m Z_i^A - P(A)\Big| > \epsilon\Big\} \le 2S(\mathcal{A},2m^2)e^{2(1+\frac{1+\rho}{2}\epsilon)}e^{-m(1+\rho)\epsilon}.$$

Similarly, Theorem 4.2.3 can be written as follows:

Theorem 4.3.2. *Assume that $\mathcal{A}$ is permissible and $s: X\to X$ is such that for some $\rho>0$, $P(A)>\epsilon$ implies $P(s^{-1}(A))>\rho\epsilon$ for all $A\in\mathcal{A}$. If $\mathbb{P}_{\mathcal{A}}=\{(p_0,p_1,p_2)\ ;\ p_2\le\gamma\}$ and $\gamma<\tilde{\epsilon}$ then*

$$\begin{aligned}\Pr\Big\{\sup_{\substack{A\in\mathcal{A}\\ \sum_{i=1}^m Z_i^A=0}} \Big|\frac{1}{2m}\sum_{i=1}^m Z_i^A - P(A)\Big| > \epsilon\Big\} &\le 2S(\mathcal{A},2m^2)\Big[\Big(1-4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}\Big)e^{-mf(m^2\rho)}\\ &\quad + 4e^{4d^2+4d}S(\mathcal{A},2m^4)e^{-2d^2m^2}e^{1+\tilde{\epsilon}}e^{-m\tilde{\epsilon}}\Big],\end{aligned}$$

where

$$\tilde{\epsilon} = \frac{1+\rho}{2}\epsilon,$$
$$f(k) = \frac{k}{m^2} + \frac{2((m\tilde{\epsilon}-1)(m-1)-k)}{m^2-k},$$

$\rho = \gamma + d$ *and* $m^2\rho \leq (m\tilde{\epsilon}-1)(m-1)$.

General Convergence Bounds

For general convergence bounds assume that there is $\rho < 2\epsilon$ so that $P(A) \leq (1+\rho)P(s^{-1}(A))$ and $P(A) \geq (1-\rho)P(s^{-1}(A))$ for all $A \in \mathcal{A}$ (we can take $\rho = (c_2 - c_1)/c_1$).

We claim that the event

$$\Big\{\sup_{A\in\mathcal{A}}\Big|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) + \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(s(X_i)) - 2P(A)\Big| > 2\epsilon\Big\}$$

is contained in

$$\Big\{\sup_{A\in\mathcal{A}}\Big|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) + \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(s(X_i)) - P(A) - P(s^{-1}(A))\Big| > 2\left(\epsilon - \rho/2\right)\Big\}.$$

Assume that $\left|\frac{1}{m}\sum\mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum\mathbf{1}_{A^*}(s(X_i)) - 2P(A^*)\right| > 2\epsilon$, then either $\frac{1}{m}\sum\mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum\mathbf{1}_{A^*}(s(X_i)) < 2P(A^*) - 2\epsilon$ or $\frac{1}{m}\sum\mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum\mathbf{1}_{A^*}(s(X_i)) > 2P(A^*) + 2\epsilon$. Using $P(A^*) \leq (1+\rho)P(s^{-1}(A^*))$ in the former and $P(A^*) \geq (1-\rho)P(s^{-1}(A^*))$ in the latter case, we get

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{A^*}(s(X_i)) < P(A^*) + P(s^{-1}(A^*)) - 2(\epsilon - \frac{\rho}{2})$$

or

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{A^*}(X_i) + \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_{A^*}(s(X_i)) > P(A^*) + P(s^{-1}(A^*)) + 2(\epsilon - \frac{\rho}{2}).$$

Thus

$$\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) + \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(s(X_i)) - P(A)1P(s^{-1}(A))\right| > 2(\epsilon - \frac{\rho}{2}).$$

That is, we can repeat the analysis of Section 4.1 with $\tilde{\epsilon} = \epsilon - \rho/2$ is place of ϵ. Then, for example, Theorem 4.1.12 translates to the following:

Theorem 4.3.3. *Let $\mathcal{A}$ be a permissible class. If $\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) \in \{0,1\}$ for all i and $s : X \to X$ is such that for some $\rho < 2\epsilon$, $(1-\rho)P(s^{-1}(A)) \le P(A) \le (1+\rho)P(s^{-1}(A))$ for all $A \in \mathcal{A}$, then*

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{2m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) + \mathbf{1}_A(s(X_i)) - P(A)\right| \ge \epsilon\right\}$$
$$\le 4e^{16(\epsilon-\rho/2)^2+16(\epsilon-\rho/2)}S(\mathcal{A}, 2m^2)e^{-8m(\epsilon-\rho/2)^2}.$$

4.4 Exponential Bounds in Learning

s:exp-b-l

Bounds of the type

$$\Pr\left\{\sup_{A\in\mathcal{A}}\left|\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}_A(X_i) - P(A)\right| \ge \epsilon\right\} \le C_1 e^{-c_2 m\epsilon^2}$$

are applied for learning an unknown target T by studying the class $\mathcal{A}\Delta T = \{A\Delta T\ ;\ A \in \mathcal{A}\}$. The bound of above type can be applied to any classes of classifiers. In particular these bounds are applied in cases where zero error classification can not be assumed and empirical error is not linked to the true error.

As an application consider the general classifier selection problem, see [18]. Let $\mathcal{C}$ be a class of decision functions of the form $\phi : \mathbb{R}^d \to \{0,1\}$. Define $\mathcal{A}$ as a collection of all sets

$$\{\{x\ ;\ \phi(x) = 1\} \times \{0\}\} \cup \{\{x\ ;\ \phi(x) = 0\} \times \{1\}\}$$

for all $\phi \in \mathcal{C}$. Using the notation $\widehat{L}_m(\phi) = 1/m\sum_{i=1}^{m}\mathbf{1}_{\{\phi(X_i)\neq Y_i\}}$ and $L(\phi) = \Pr\{\phi(X) = Y\}$ we study

$$\Pr\{L(\phi_m^*) - \inf_{\phi\in\mathcal{C}} L(\phi) > \epsilon\},$$

where ϕ_m^* denotes the classifier minimizing $\widehat{L}_m(\phi)$ over the class $\mathcal{C}$. This can be bounded by the uniform convergence as

$$L(\phi_m^*) - \inf_{\phi\in\mathcal{C}} L(\phi) \le 2\sup_{\phi\in\mathcal{C}} \left|\widehat{L}_m(\phi) - L(\phi)\right|.$$

To modify the problem for side information assume that we observe $\{(x_1, y_1), (s(x_1), y_1^s), \ldots, (x_m, y_m), (s(x_m), y_m^s)\}$. Let

$$\hat{\hat{L}}_m(\phi) = \frac{1}{2m}\sum_{i=1}^{m} \mathbf{1}_{\{\phi(X_i)\neq Y_i\}} + \mathbf{1}_{\{\phi(s(X_i))\neq Y_i^s\}}$$

and let ϕ_m^* minimize $\hat{\hat{L}}_m$. Then also

$$L(\phi_m^*) - \inf_{\phi\in\mathcal{C}} L(\phi) \le 2\sup_{\phi\in\mathcal{C}} \left|\hat{\hat{L}}_m(\phi) - L(\phi)\right|$$

and we can relate the problem to convergence bounds studied in Section 4.1.

Next we briefly illustrate convergence bounds with consistent algorithms (with respect to the original data and the side information) with measure preserving side information. The following proposition is an example to show that in some cases we can split the learning bound in two parts: one in which $p_2 = 0$ and one (with small probability) in which we apply the bound with exponential term $e^{-m\epsilon}$.

Proposition 4.4.1. *Let X be a metric space and let* $\mathrm{diam}(A) = \sup_{x,y\in A} d(x,y)$. *Consider the concept class $\mathcal{C} = \{A \subseteq X\ ;\ \mathrm{diam}(A) \le C, P(A) \ge \lambda\}$ and side information mapping s such that $d(x, s(x)) > 2C$ for all $x \in X$. Then for any $T \in \mathcal{C}$ and any consistent estimate $\widetilde{h}_m$ taking values in $\mathcal{C}$, with probability at least, $(1-(1-\lambda)^m)$, $p_2 = \Pr\{\mathbf{1}_{\widetilde{h}_m\Delta T}(X_i) + \mathbf{1}_{\widetilde{h}_m\Delta T}(s(X_i)) = 2\} = 0$. Further,*

e:ex-learn

$$\sup_{T\in\mathcal{C}} P^m\{\bar{x} \in X^m\ ;\ \exists\,\widetilde{h}_m \text{ consistent}, d_P(T, \widetilde{h}_m) > \epsilon\}$$
$$\le 2\left(\frac{2em^2}{d}\right)^d \left[(1-(1-\lambda)^m)e^{-2m\epsilon}e^{2+2\epsilon} + (1-\lambda)^m e^{-m\epsilon}e^{1+\epsilon}\right], \quad (4.4.9)$$

where d is the VC-dimension of $\mathcal{C}$.

Proof. After a positive sample has been obtained, $\widetilde{h}_m$ and T intersect each other and thus $\widetilde{h}_m \Delta T$ is contained in a set with diameter at most $2C$. As $d(x, s(x)) > 2C$, it is not possible that both x and $s(x)$ are in $\widetilde{h}_m \Delta T$.

Let $A_m = \widetilde{h}_m \Delta T$ and $\mathcal{A} = \mathcal{C} \Delta T$. Then the LHS of (4.4.9) is bounded by

$$\sup_{T \in \mathcal{C}} \Pr\{ \sup_{\substack{A_m \in \mathcal{A} \\ \sum \mathbf{1}_{A_m}(X_i) + \mathbf{1}_{A_m}(s(X_i)) = 0}} \left| \frac{1}{2m} \sum_{i=1}^{m} \mathbf{1}_{A_m}(X_i) + \mathbf{1}_{A_m}(s(X_i)) - P(A_m) \right| > \epsilon \},$$

which is in turn bounded using Lemma 4.2.1 to obtain

$$\sup_{T \in \mathcal{C}} 2S(\mathcal{A}, 2m^2) E \left[e^{\frac{-k^{A_m}}{m}} e^{\frac{-ml^{A_m}}{m^2 - k^{A_m}}} \right].$$

By using Sauer's lemma to bound $S(\mathcal{A}, 2m^2)$ and the fact that $k^{A_m} = 0$ with probability at least $(1 - (1 - \lambda)^m)$ we obtain the bound

$$2 \left(\frac{2em^2}{d} \right)^d \left[(1 - (1 - \lambda)^m) \, e^{-2m\epsilon} e^{2+2\epsilon} + (1 - \lambda)^m e^{-m\epsilon} e^{1+\epsilon} \right].$$

$\square$

4.5 Open Problems with Side Information as a Complexity Constraint

We propose a possible framework for incorporating side information into learning analysis when original samples and side information samples are fundamentally different objects. A measure preserving side information avoided the problem as it was possible to treat the side information samples in a way similar to the original samples. In the following approach the idea is to capture side information as a restrictive condition for consistency and thus treat the original data and the side information data differently. In the learning situation in some sense the side information has to be treated differently from the original samples, because at the end we want to classify correctly a future unseen original sample and the generalization ability with respect to side information samples is irrelevant.

This section is organized as follows: we begin by introducing shattering with side information and we show that it has nice VC-dimension like properties. In Section 4.5.2 we show that our proposed complexity dimension has the property of "probable smoothness" and it fits into the framework introduced in [34]. We finish with examples of side information shattering in Section 4.5.3 and open problems in Section 4.5.4.

4.5.1 Side Information Shattering

Let $\mathcal{A}$ be a concept class, $T \in \mathcal{A}$, let $s\colon X \to X$ be the side information mapping, and assume that $\{x_1, \dots, x_m\} \cap \{s(x_1), \dots, s(x_m)\} = \emptyset$ with probability 1.

Definition 7. Define

$$\begin{aligned} N_{\mathcal{A},s,T}(x_1, \dots, x_m) &= \Big|\Big\{(1_A(x_1), \dots, 1_A(x_m)) \ ; \ A \in \mathcal{A} \text{ and} \\ &\qquad\qquad 1_A(s(x_i)) = 1_T(s(x_i)), i = 1, \dots, m\Big\}\Big| \\ &= \Big|\Big\{(1_A(x_1), \dots, 1_A(x_m)) \ ; \ A \in \mathcal{A}_{s,T}(x_1, \dots, x_m)\Big\}\Big|, \end{aligned}$$

$$\begin{aligned} S_T(\mathcal{A}, s, m) &= \max_{(x_1, \dots, x_m)} N_{\mathcal{A},s,T}(x_1, \dots, x_m), \\ S(\mathcal{A}, s, m) &= \max_{T \in \mathcal{A}} S_T(\mathcal{A}, s, m). \end{aligned}$$

We say that $(x_1, \dots, x_m)$ is *shattered by* $(\mathcal{A}, s, T)$ if $N_{\mathcal{A},s,T}(x_1, \dots, x_m) = 2^m$, $\mathrm{VC}_{s,T}(\mathcal{A})$ is the largest m such that there exists $(x_1, \dots, x_m)$ that is shattered by $(\mathcal{A}, s, T)$, and $\mathrm{VC}_s(\mathcal{A}) = \max_{T \in \mathcal{A}} \mathrm{VC}_{s,T}(\mathcal{A})$. We call $\mathrm{VC}_s(\mathcal{A})$ as the *SI-dimension of* $\mathcal{A}$.

We begin by showing that both $\mathrm{VC}_{s,T}(\mathcal{A})$ and $\mathrm{VC}_s(\mathcal{A})$ have nice VC-dimension like properties: taking symmetric differences does not change the dimension and

Sauer's lemma holds. The proofs for side information can be related to VC-dimension results.

l:symdiff

Lemma 4.5.1. *For $T \in \mathcal{A}$, $VC_{s,T}(\mathcal{A}) = VC_{s,\emptyset}(\mathcal{A}\Delta T)$, where $\mathcal{A}\Delta T = \{A\Delta T\ ;\ A \in \mathcal{A}\}$.*

Proof. One can show that if $G \subseteq \{x_1, \ldots, x_m\}$ the mapping $B \mapsto B\Delta(G \cap T)$ is a bijection between

$$\{G \cap A\ ;\ A \in \mathcal{A} \text{ and } 1_A(s(x_i)) = 1_T(s(x_i)),\ i = 1, \ldots, m\},$$

and

$$\begin{aligned}&\{G \cap (A\Delta T)\ ;\ A\Delta T \in \mathcal{A}\Delta T \text{ and } 1_{A\Delta T}(s(x_i)) = 0,\ i = 1, \ldots, m\}\\ &\qquad = \{G \cap (A\Delta T)\ ;\ A\Delta T \in \mathcal{A}\Delta T \text{ and } 1_A(s(x_i)) = 1_T(s(x_i)),\ i = 1, \ldots, m\}.\end{aligned}$$

□

l:si-sauer

Lemma 4.5.2 (Sauer's lemma).

$$N_{\mathcal{A},s,T}(x_1, \ldots, x_m) \le \sum_{i=0}^{\mathrm{VC}_{s,T}(\mathcal{A})} \binom{m}{i} \qquad \textit{for all } (x_1, \ldots, x_m).$$

Proof. Suppose that $\mathrm{VC}_{s,T}(\mathcal{A}) = d$. Pick any sequence $(x_1, \ldots, x_m)$ with $m > d$. $N_{\mathcal{A},s,T}(x_1, \ldots, x_m) = |\{(\mathbf{1}_A(x_1), \ldots, \mathbf{1}_A(x_m))\ ;\ A \in \mathcal{A}_{s,T}(x_1, \ldots, x_m)| \le |\{(\mathbf{1}_A(x_1), \ldots, \mathbf{1}_A(x_m))\ ;\ A \in \mathcal{A}_{s,T}(x_1, \ldots, x_d)|$. Note that the VC-dimension of $\mathcal{A}_{s,T}(x_1, \ldots, x_d)$ is also d and by Sauer's lemma

$$|\{(\mathbf{1}_A(x_1), \ldots, \mathbf{1}_A(x_m))\ ;\ A \in \mathcal{A}_{s,T}(x_1, \ldots, x_d)| \le \sum_{i=1}^{d} \binom{m}{i}.$$

Hence

$$N_{\mathcal{A},s,T}(x_1, \ldots, x_m) \le \sum_{i=0}^{\mathrm{VC}_{s,T}(\mathcal{A})} \binom{m}{i} \qquad \text{for all } (x_1, \ldots, x_m).$$

□

Typically, we write

$$N_{\mathcal{A},s,T}(x_1,\ldots,x_m) \leq \sum_{i=0}^{\mathrm{VC}_{s,T}(\mathcal{A})} \binom{m}{i} \leq \left(\frac{em}{\mathrm{VC}_{s,T}(\mathcal{A})}\right)^{\mathrm{VC}_{s,T}(\mathcal{A})}$$

and also

$$S(\mathcal{A},s,m) \leq \left(\frac{em}{\mathrm{VC}_s(\mathcal{A})}\right)^{\mathrm{VC}_s(\mathcal{A})}$$

4.5.2 SI-Dimension in Learning

s:si-dim

The setting of structural risk minimization over data-dependent hierarchies by Shawe-Taylor et. al. [34] allows working with luckiness (or unluckiness) functions that behave in the learning bounds much like the ordinary VC-dimension does. We can show that SI-dimension, and even the data dependent version of that, can be used as a luckiness function (actually an unluckiness function) and both have the property of "probabable smoothness". Note that the framework of structural risk minimization is rather general and it subsumes the PAC learning setting.

Let $L\colon X^m \times \mathcal{H} \to \mathbb{R}^+$ be the luckiness of a particular hypothesis $h \in \mathcal{H}$ with respect to training samples. Similarly, let $U\colon X^m \times \mathcal{H} \to \mathbb{R}^+$ be the unluckiness function. One can work with either L or U.

Define the level of a function $h \in \mathcal{H}$ relative to L and x (or U and x) by the function

$$l(x,h) = \left|\left\{b \in \{0,1\}^m \;;\; \exists g \in \mathcal{H}, g(x) = b, L(x,g) \geq L(x,h)\right\}\right|$$

or

$$l(x,h) = \left|\left\{b \in \{0,1\}^m \;;\; \exists g \in \mathcal{H}, g(x) = b, U(x,g) \leq U(x,h)\right\}\right|.$$

If $x, y \in X^m$ let $xy = (x_1,\ldots,x_m,y_1,\ldots,y_m)$. An α-subsequence of a vector x is a vector x' obtained from x by deleting a fraction of at most α coordinates.

Definition 8 (Probable smoothness of luckiness functions, [34]). A luckiness function $L(x,h)$ (or unluckiness $U(x,h)$) defined on a function class $\mathcal{H}$ is *probably smooth* with respect to functions $\eta(m,L,\delta)$ and $\phi(m,L,\delta)$, if, for all targets $T \in \mathcal{H}$ and for every distribution P

$$P^{2m}\big\{xy\ ;\ \exists h \in \mathcal{H}, Er_x(h) = 0, \forall x'y' \subseteq_\eta xy, l(x'y',h) > \phi(m, L(x,h), \delta)\big\} \le \delta,$$

where $\eta = \eta(m, L(x,h), \delta)$ and $Er_x(h) = |\{i\ ;\ h(x_i) \neq T(x_i)\}|$ is the number of errors made by h in the sample x.

Next we show that SI-dimension can be used as an unluckiness function.

l:lemma3

Lemma 4.5.3. *$U(x,A) = VC_s(\mathcal{A})$ for all $x \in X^m$ and all $A \in \mathcal{A}$ is probably smooth with respect to $\phi(m,U,\delta) = (2em/U)^U$ and $\eta(m,U,\delta) = 0$.*

Proof. We need to show that

e:si-sau

$$P^{2m}\big\{xy\ ;\ \exists A \in \mathcal{A}, Er_x(A) = 0, l(xy,A) > \phi(m,U,\delta)\big\} \le \delta. \tag{4.4.10}$$

But

$$\begin{aligned} l(xy,A) &= \Big|\big\{b \in \{0,1\}^{2m}\ ;\ \exists G \in \mathcal{A}, G(xy) = b, U(xy,G) \le U(xy,A)\big\}\Big| \\ &= \text{number of } 0,1\text{-assignments with } G \in \mathcal{A} \text{ and with side information} \\ &= S(\mathcal{A}, s, 2m) \\ &\le \left(\frac{2em}{\mathrm{VC}_s(\mathcal{A})}\right)^{\mathrm{VC}_s(\mathcal{A})} = \phi(m,U,\delta) \end{aligned}$$

by Lemma 4.5.2. Hence (4.4.10) is true with any $\delta > 0$. □

We can do even more in this framework: Let $\mathcal{H}(x_1,\ldots,x_m)$ be the concepts restricted to $x_1,\ldots,x_m$ and consistent with side information $s(x_1),\ldots,s(x_m)$. Define $U(x,h) = \mathrm{VC}_s(\mathcal{H}(x_1,\ldots,x_m)) = \sup_T \mathrm{VC}_{s,T}(\mathcal{H}(x_1,\ldots,x_m))$.

l:datadep

Lemma 4.5.4. *$U(x,h)$ defined above is probably smooth with respect to $\eta(m,U,\delta) = 0$ and $\Phi(m,u,\delta) = (2em/\alpha U)^{\alpha U}$, where $\alpha = \alpha(U,\delta) = 3.08(1+\frac{1}{U}\log\frac{1}{\delta})$.*

Sketch of proof. Shawe-Taylor et. al. [34] prove that $U(x,h) = VC\dim(\mathcal{H}|_x)$ is probably smooth with respect to $\eta(m,U,\delta)$ and $\Phi(m,U,\delta)$ defined above. The idea is to show that with high probability the ratio between the VC-dimensions obtained by restricting $\mathcal{H}$ to single and double samples is not greater than α. Their methods work for proving this lemma, too. We need that whenever $z_1,\ldots,z_d$ from the double sample are shattered by $\mathcal{H}(x_1,\ldots,x_m,y_1,\ldots,y_m)$, then also any samples of those d that are in the "x-part" are shattered by $\mathcal{H}(x_1,\ldots,x_m)$. This is true as $\mathcal{H}(x_1,\ldots,x_m,y_1,\ldots,y_m) \subseteq \mathcal{H}(x_1,\ldots,x_m)$. □

4.5.3 Examples of SI-dimension

s:si-dim-ex

In this section we give some examples of SI-dimension to illustrate the effect of adding consistency contrains. It turns out the in the simple examples analyzed the restrictive power of side information is not very strong. As we have earlier seen the benefit of the side information depends on the probability distribution, the side information mapping and the target concept. How much smaller the SI-dimension is in comparison to the VC-dimension depends on how "special" the VC-shattered sets are and if the side information mapping is "compatible" with the shattered sets.

ex:si-conv

Example 4.5.5. Let $\mathcal{A}$ be the collection of convex sets in $\mathbb{R}^2$, consider any target $T \in \mathcal{A}$ and let $s(re^{i\theta}) = re^{i(\theta+\lambda)}$, $\lambda \neq 0$. $\mathrm{VC}(\mathcal{A}) = \infty$, because any number of points on the unit sphere can be shattered. Similarly, $\mathrm{VC}_s(\mathcal{A}) = \infty$. Note that here points $x_1,\ldots,x_m$ and $s(x_1),\ldots,s(x_m)$ mix well, but the side information mapping preserves the special structure of shattered sets. For other side information mappings one expects improvement in the SI-dimension if the points $x_1,\ldots,x_m$ and $s(x_1),\ldots,s(x_m)$ are not separated.

Example 4.5.6. Let $\mathcal{A}$ = open intervals $(a,b) \subset (0,1)$. Pick $x_1, x_2 \in (0,1)$, $x_1 < x_2$. If $1_T(s(x_i)) = 1_{[x_1,x_2]}(x_i)$, $i = 1,2$, then $\mathrm{VC}_{s,T}(\mathcal{A}) = 2$, otherwise $\mathrm{VC}_{s,T}(\mathcal{A}) = 1$.

Next we discuss SI-dimension for the collection of half-spaces through origin in $\mathbb{R}^k$. We show first that if there exists a point with fixed sign assigment, one can shatter only $k-1$ points, while being consistent with the fixed assigment. A surprising result is that adding more points with fixed sign assigments may not reduce the shattering ability further.

l:lemma5

Lemma 4.5.7. *Let G be a finite dimensional vector space of real functions on $\mathbb{R}^d$. Then the class of sets*

$$\mathcal{A} = \big\{\{x \; ; \; g(x) \geq 0\} \; ; \; g \in G\big\}$$

satisfies $VC_{s,T}(\mathcal{A}) \leq r-1$ for any T and any s, where $r = \dim(G)$.

Proof. We show that it is not possible to shatter r points if there exists one point with fixed assignment. (Here it is convenient to work with $+, -$ assignments.) Take arbitrary points $x_1, x_2, \ldots, x_r$ and a side information point x_{r+1} with assignment $+$ ($-$ will be done in a similar way).

Define a linear mapping $L\colon G \to \mathbb{R}^{r+1}$ by

$$L(g) = \{g(x_1), g(x_2), \ldots, g(x_r), g(x_{r+1})\}.$$

As $\dim L(G) = r$ there exists $(\gamma_1, \ldots, \gamma_r, \gamma_{r+1}) \neq (0, \ldots, 0, 0)$ such that

e:lemma5-1

$$\gamma_1 g(x_1) + \cdots + \gamma_r g(x_r) + \gamma_{r+1} g(x_{r+1}) = 0 \qquad \forall g \in G. \tag{4.4.11}$$

If necessary, multiply by -1 so that $\gamma_{r+1} \geq 0$ and write (4.4.11) as

e:lemma5-2

$$\sum_{i \; ; \; \gamma_i \geq 0} \gamma_i g(x_i) = - \sum_{i \; ; \; \gamma_i < 0} \gamma_i g(x_i). \tag{4.4.12}$$

Suppose that there exists $g \in G$ such that the set $\{x \; ; \; g(x) \geq 0\}$ picks only the x_i's on the left-hand side of (4.4.12). Then the left-hand side is non-negative whereas the right-hand side of (4.4.12) is negative. Thus the set of r points can not be side information shattered.

Now if $g(x_{r+1})$ has $-$ assignment, if necessary, multiply by -1 so that $\gamma_{r+1} \leq 0$. The rest of the proof is the same. $\square$

We illustrate by two examples that having $k-1$ linearly independent points with fixed sign assigments does not reduce the shattering power. This is because given the side information mapping $s(x)$, SI-shattering allows us to design not only the shattered points but also it gives the freedom to choose a target T. For next two propositions we emphasize vectors to distinguish them from components of a vector.

Proposition 4.5.8. *Let $\mathcal{A}$ be the collection of half-spaces through origin in $\mathbb{R}^k$. Then there exists a side information mapping $s : \mathbb{R}^k \to \mathbb{R}^k$ such that $\mathcal{A}$ can side information shatter $k-1$ points in $\mathbb{R}^k$.*

Proof. Let $s(\vec{x}_i) = \vec{e}_i$ for $i = 1, \ldots, k-1$ and without loss of generality we assume that the sign assignment to each $s(\vec{x}_i)$ is $+$.

Let

$$\begin{array}{rcl} \vec{x}_1 & = & (1 \quad -1 \quad 0 \ \ldots \ 0) \\ \vec{x}_2 & = & (0 \quad 1 \quad -1 \ \ldots \ 0) \\ \vdots & & \quad \vdots \\ \vec{x}_{k-2} & = & (0 \ \ldots \ 1 \quad -1 \quad 0) \\ \vec{x}_{k-1} & = & (0 \quad 0 \ \ldots \ 0 \quad 1). \end{array}$$

We consider first sign assignments to $\vec{x}_1, \ldots, \vec{x}_{k-2}$ and we can think $\vec{x}_i \in \mathbb{R}^{k-1}$ for $i = 1, \ldots, k-2$. Given a sign assignment to $\vec{x}_1, \ldots, \vec{x}_{k-2}$, we need to find a vector $\vec{c} \in \mathbb{R}^{k-1}$ such that $\text{sign}(\vec{c} \cdot \vec{x}_1, \ldots, \vec{c} \cdot \vec{x}_{k-2}) = (s_1, \ldots, s_{k-2})$, where we use $\text{sign}(y) = +1$, if $y \geq 0$, and -1 otherwise , together with $\text{sign}(y_1, \ldots, y_{k-2}) = (\text{sign}(y_1), \ldots, \text{sign}(y_{k-2}))$. Positive sign assigments for side information is equivalent for the $k-1$ first components of $\vec{c}$ being positive.

Let $\vec{s} = (s_1, \ldots, s_{k-2})$ be a sign assignment to $(\vec{x}_1, \ldots, \vec{x}_{k-2})$. We build recursively a vector $\vec{c} \in \mathbb{R}^{k-1}$ that agrees with sign assignment $\vec{s}$ and is consistent with the side information constraint. Let $c_1 = 1$, where c_1 is the first component of $\vec{c}$ and then recursively $c_{i+1} = 2^{-s_i} c_i$ for $i = 1, \ldots, k-2$. Now $\text{sign}(\vec{c} \cdot \vec{x}_1, \ldots, \vec{c} \cdot \vec{x}_{k-2}) =$

$\text{sign}(1 - 2^{-s_i}, \ldots, 1 - 2^{-s_{k-2}}) = (s_1, \ldots, s_{k-2})$. Thus, when solving the problem in $\mathbb{R}^{k-1}$, all possible sign assignments can be obtained to $(\vec{x}_1, \ldots, \vec{x}_{k-2})$.

Any sign assignment to $(\vec{x}_1, \ldots, \vec{x}_{k-2}, \vec{x}_{k-1})$ can be obtained with vector $(\vec{c}, c') \in \mathbb{R}^k$, where $\vec{c} \in \mathbb{R}^{k-1}$ is the vector constructed above and $c' \in \mathbb{R}$ is positive if $\text{sign}(\vec{x}_{k-1}) = 1$ and negative otherwise. Note that taking the kth component of the vector to be negative does not violate the side information condition.

If the sign assigment to $s(\vec{x}_i)$ were different we can change signs in the columns of $\vec{x}_i, \ldots, \vec{x}_{k-2}$ and modify the recursive definition to give correct sign for the components of $\vec{c}$. □

p:sish-half

Proposition 4.5.9. *Let $\mathcal{A}$ be the collection of half-spaces through origin in $\mathbb{R}^k$ and let the side information mapping to be $s(x_1, \ldots, x_k) = (0, x_1, \ldots, x_{k-1})$. Then there is a target T such that $VC_{s,T}(\mathcal{A}) = k - 1$.*

Proof. Consider points $\vec{x}_1 = (1, 0, 0, \ldots, 0, -1)$, $\vec{x}_2 = (0, 1, 0, \ldots, -1), \ldots, \vec{x}_{k-1} = (0, 0, \ldots, 1, -1)$. Then $s(\vec{x}_i) = \vec{e}_{i+1}$ and let all side information points have sign assigment $+$. Given a sign assignment $\vec{s} = (s_1, \ldots, s_k)$ to $(\vec{x}_1, \ldots, \vec{x}_k)$ we form a classifying vector $\vec{c}$ so that $c_k = 2$ and $c_i = c_k + s_i$ for $i = 1, \ldots, k-1$. This is consistent with the side information constraint and $\text{sign}(\vec{c} \cdot \vec{x}_1, \ldots, \vec{c} \cdot \vec{x}_k) = \text{sign}(c_1 - c_k, \ldots, c_{k-1} - c_k) = (s_1, \ldots, s_k)$. □

4.5.4 Open Problems

s:si-dim-o

Our motivation to study the problem of side information was the control system application introduced in Chapter 3. To simplify the problem we assume that only the output preceeding the final output, i.e., $\text{sign}(y_{k-1})$, is available and thus the side information mapping is $s(u_1, \ldots, u_k) = (0, u_1, \ldots, u_{k-1})$. (One expects that the value of additional outputs decreases as the effective length of the input decreases, i.e., input $(0, \ldots, 0, u_1)$ is likely to provide less information than the input

$(0, u_1, \ldots, u_{k-1})$.)

Proposition 4.5.9 shows that the side information mapping $s(u_1, \ldots, u_k) = (0, u_1, \ldots, u_{k-1})$ does not have strong restrictive power for half-spaces. This is because the side information constraint is in a lower dimension and, in general, half-spaces in $\mathbb{R}^k$ can shatter in the VC-sense all affinely independent $(k+1)$-element subsets. That is, for half-spaces the shattered sets are by no means special and adding side information as a restriction for shattering may not be strong.

Proposition 4.5.9 does not serve as an example for control systems as the vector $\vec{c}$ constructed to realize a given sign assignment is far from being n-recursive. For control systems the restrictive power of side information is an open problem. So far we have only upper and lower bounds for the VC-dimension associated to control systems. One should aim at studying how special the shattered sets are. The following approach to make precise the genericity of shattered sets was introduced by Sontag, see [11, 36]: Fix a concept class and assume that the inputs are k-vectors. For each ℓ let S_ℓ be the subset of $\mathbb{R}^{\ell k}$ consisting of the ordered sets of ℓ inputs that are shattered. Then define as a new notation of dimension

$$\mu := \sup\{\ell \geq 1\ ;\ S_\ell \text{ is a dense subset of } \mathbb{R}^{\ell k}\}.$$

However, even if shattered sets are "special", we have seen in Example 4.5.5 that side information may not reduce the shattering power, if the side information mapping maintains the special structure of shattered sets.

In side information problems the advantage of side information depends on the learning algorithm, the distribution of the samples and also on the target concept. Given a side information mapping, SI-shattering allows one the choose the worst possible target, and inevitably SI-shattering becomes worst case analysis. One possible way to overcome the worst case analysis is to consider $\mathrm{VC}_{s,T}(\mathcal{H}(x_1, \ldots, x_m))$, where $\mathcal{H}(x_1, \ldots, x_m)$ in the class of concepts restricted to $x_1, \ldots, x_m$ and consistent with the side information $s(x_1), \ldots, s(x_m)$. (We indicated that $\sup_T \mathrm{VC}_{s,T}(\mathcal{H}(x_1, \ldots, x_m))$

can be used as an unluckiness function, in the same way the quantity without supremum over the targets can be used an unluckiness function.) Shawe-Taylor et. al. [34] motivate the data-dependent VC-dimension by refering to results showing that classes with infinite VC-dimension may still be learnable provided that the distribution of inputs is concentrated on the regions of the input space where the set of hypothesis has low VC-dimension or that in some cases shattered set are "special". Hence, by using data-dependent VC-dimension as an unluckiness function, it is possible to learn succesfully if a region with low VC-dimension is witnessed. In addition to above, our motivation is to be able to observe if the side information mapping and the target are "compatible".

Appendix A

Example of the Goldberg-Jerrum Bound

a:gold-ex

The purpose of the following example is to illustrate the function $y = \mathbf{P}/\mathbf{Q}$ used in the Goldberg-Jerrum technique together with the sequence of polynomial evaluations involved and a table for the final output depending on the outcomes of the polynomial evaluations.

Take $m = 1$, $n = 2$, $k = 2$, and assume that A has complex eigenvalues $a \pm ib$. Take control functions to be $\omega_1(t) = e^t$ and $\omega_2(t) = e^{2t}$. Then

$$y(1) = \sum_{l=1}^{2} \alpha_l \sum_{j=1}^{2} g_j \int_0^1 \xi_l(t)\omega_j(t)\, dt,$$

where $\xi_1(t) = e^{at}\sin(bt)$, $\xi_2(t) = e^{at}\cos(bt)$, α_1, α_2, a, b, $e^a \sin b$ and $e^a \cos b$ are system parameters and g_1, g_2 are input parameters.

By using formulas

$$\int_0^1 e^{\tilde{a}t}\sin(\tilde{b}t)\, dt = \frac{e^{\tilde{a}}(\tilde{a}\sin\tilde{b} - \tilde{b}\cos\tilde{b}) + \tilde{b}}{\tilde{a}^2 + \tilde{b}^2}, \quad \text{and}$$
$$\int_0^1 e^{\tilde{a}t}\cos(\tilde{b}t)\, dt = \frac{e^{\tilde{a}}(\tilde{a}\cos\tilde{b} + \tilde{b}\sin\tilde{b}) - \tilde{a}}{\tilde{a}^2 + \tilde{b}^2},$$

we calculate the integrals appearing in the rationality condition, and we call them $r_{11}, \ldots, r_{22}$:

$$\int_0^1 \xi_1(t)\omega_1(t)\,dt = \int_0^1 e^{(a+1)t}\sin(bt)\,dt = \begin{cases} r_{11}, & \text{if } (a+1)^2+b^2 \neq 0, \\ 0, & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$

$$\int_0^1 \xi_1(t)\omega_2(t)\,dt = \begin{cases} r_{12}, & \text{if } (a+2)^2+b^2 \neq 0, \\ 0, & \text{if } (a+2)^2+b^2 = 0, \end{cases}$$

$$\int_0^1 \xi_2(t)\omega_1(t)\,dt = \begin{cases} r_{21}, & \text{if } (a+1)^2+b^2 \neq 0, \\ 1, & \text{if } (a+1)^2+b^2 = 0, \end{cases}$$

$$\int_0^1 \xi_2(t)\omega_2(t)\,dt = \begin{cases} r_{22}, & \text{if } (a+2)^2+b^2 \neq 0, \\ 1, & \text{if } (a+2)^2+b^2 = 0. \end{cases}$$

The computation of

$$\operatorname{sign} y(1) = \operatorname{sign}\left(\sum_{l=1}^{2}\alpha_l \sum_{j=1}^{2} g_j \int_0^1 \xi_l(t)\omega_j(t)\,dt\right)$$

is divided into three cases:

- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 \neq 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_1 g_2 r_{12} + \alpha_2 g_1 r_{21} + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_1}{Q_1}\right).$$

- Case $(a+1)^2+b^2 = 0$, $(a+2)^2+b^2 \neq 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_2 r_{12} + \alpha_2 g_1 + \alpha_2 g_2 r_{22}) = \operatorname{sign}\left(\frac{P_2}{Q_2}\right).$$

- Case $(a+1)^2+b^2 \neq 0$, $(a+2)^2+b^2 = 0$:

$$\operatorname{sign} y(1) = \operatorname{sign}(\alpha_1 g_1 r_{11} + \alpha_2 g_2 r_{21} + \alpha_2 g_2) = \operatorname{sign}\left(\frac{P_3}{Q_3}\right).$$

Thus we have three different expressions of the form $\dfrac{\mathbf{P}}{\mathbf{Q}}$.

Next we form the Boolean formula, $F = \operatorname{sign} y(1)$, evaluating polynomials $f_1 = (a+1)^2 + b^2 = 0$, $f_2 = (a+2)^2 + b^2 = 0$, $P_i > 0$, $Q_i > 0$, for $i \in \{1, 2, 3\}$. In the following table 1 means true and 0 means false for the above polynomial evaluation ($* = 1$ or 0, i.e., extend the table).

$f_1 = 0$	$f_2 = 0$	$P_1 > 0$	$Q_1 > 0$	$P_2 > 0$	$Q_2 > 0$	$P_3 > 0$	$Q_3 > 0$	F
0	0	1	1	$*$	$*$	$*$	$*$	1
0	0	1	0	$*$	$*$	$*$	$*$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
1	0	$*$	$*$	1	1	$*$	$*$	1
1	0	$*$	$*$	1	0	$*$	$*$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
0	1	$*$	$*$	$*$	$*$	0	0	1

In this case (see the statement of Goldberg-Jerrum bounds) $\Lambda = \{\alpha_1, \alpha_2, a, b, e^a \cos b, e^a \sin b\}$, $X = \{g_1, g_2\}$, $s = 8$, $d = 12$, and $l = 6$.

Appendix B

Examples of the Simulation Programs

a:numres-app

We provide two Matlab programs that we used to carry out numerical simulations. The algorithm described in Section 3.4 produces naturally two $N \times M$ matrices, called `err` and `si_err` in the code, containing the misclassification errors without and with side information. Each row corresponds to one experiment of sample length M and the mth column is the error for sample m in each experiment. In simulation we form the matrices by columns for two reasons. First, in this way we utilize fast vector operations in Matlab instead of slow for-loops, which speeds the simulation significantly. Second, this code is easy to modify to utilize cell matrices; `err` and `si_err` are now vectors of length M and each entry is a vector of length N containing misclassification errors in each repeated experiment. In this way the simulation can be carried out with the student version of Matlab in which the matrices can have at most 16384 elements.

The following programs are used to simulate density $2t$. The complete simulation in Matlab consists of giving the beginning and the endpoint of the interval in variables `aa` and `bb` and then running `simu2t` and `prob`.

```
% File: simu2t.m
% learning rate simulation with sample drawn from density 2 t
% algorithm: smallest interval that is consistent with data
% produces matrices:
% err    learning errors for each sample (column coordinate),
%        repeated simulations (row coordinate)
% si_err same as err, but now the algorithm takes advantage on side
%        information as well

% initialize:

ssize = 70;
repeat = 16384;

si_err = cell(ssize,1);
err = cell(ssize,1);
for i=1:ssize
si_err{i} = NaN * ones(repeat,1);
err{i} = NaN * ones(repeat,1);
end

% (aa,bb) is the target interval, aa and bb defined outside this program

a = aa * ones(repeat,1);
b =  bb * ones(repeat,1);

x_min = b;
x_max = a;
si_x_min = b;
si_x_max = a;

% simulation:
% vectorized fast code, err, and si_err matrices are generated by columns
% distribution 2t is obtained by taking sample from uniform and taking
% sqrt from that.
% x is the current sample and si_x is the  side info translate
% 1) find if current sample is between endpoints of the target and
% the endpoints of current estimate, if so update the estimate
% 2) calculate the error between the target and the new estimate

for m=1:ssize,
    x_tmp =rand(repeat,1);
    x = sqrt(x_tmp);
    si_x = x + 0.5 - floor(x + 0.5);

    ind = find( a < x & x < x_min);
    x_min(ind) = x(ind);

    ind = find (x_max < x & x < b);
    x_max(ind) = x(ind);

% error calculation w.r.t. density 2 t
```

```
    ind = find( x_min == b | x_max == a);
    if (~isempty(ind))
        err{m}(ind) = b(ind).^2  - a(ind).^2;
    end

    ind = find(~( x_min == b | x_max == a));
    if (~isempty(ind))
        err{m}(ind) =  x_min(ind).^2 - a(ind).^2 + ...
                       b(ind).^2 - x_max(ind).^2;
    end

% the same with side information si_x

    ind = find( a < si_x & si_x < si_x_min);
    si_x_min(ind) = si_x(ind);

    ind = find (si_x_max < si_x & si_x < b);
    si_x_max(ind) = si_x(ind);

    % calculating error  w.r.t, density 2 t
    a_esti = min(si_x_min, x_min);
    b_esti = max(si_x_max, x_max);
    ind = find( a_esti == b | b_esti == a);
    if (~isempty(ind))
        si_err{m}(ind) = b(ind).^2 -a(ind).^2;
    end

    ind = find(~( a_esti == b | b_esti == a));
    if (~isempty(ind))
        si_err{m}(ind) = a_esti(ind).^2 -a(ind).^2 + ...
                         b(ind).^2- b_esti(ind).^2;
    end
end
```

```
% File: prob.m
% given matrices err and si_err prob.m calculates the procentage of
% simulations that have learning error e or larger (relaxing for 100
% machine epsilons.
% procuces vectors: pr_err and si_pr_err that give the error
% probabilities for each sample size

e = 0.1
M = size(err{1},1);
for m=1:length(err)
    ind1 = find (err{m} > e * ones(M,1) + 100*eps);
    ind2 = find (si_err{m} > e * ones(M,1) + 100*eps);
    pr_err(m) = length(ind1) / M;
    si_pr_err(m) = length(ind2) / M;
end
```

Bibliography

[1] Y. Abu-Mostafa, *Hints and the VC dimension*, Neural Computation **5** (1993), 278–288.

[2] ———, *Learning from hints*, Journal of Complexity **10** (1994), 165–178.

[3] ———, *Hints*, Neural Computation **7** (1995), 639–671.

[4] N. Alon, S. Ben-David, N Cesa-Bianchi, and D. Haussler, *Scale-sensitive dimensions, uniform convergence and learnability*, Journal of the ACM **44** (1997), no. 4, 615–631.

[5] M. Anthony and B. Bartlett, *Learning in neural networks: theoretical foundations*, to appear.

[6] M Anthony and N. Biggs, *Computational learning theory*, Cambridge University Press, 1992.

[7] N. Balakrishnan and A. Clifford Cohen, *Order statistics and inference: estimation methods*, Academic Press, 1991.

[8] P. Bartlett, P. Long, and R. Williamson, *Fat-shattering and learnability of real-valued functions*, Proceedings of the Seventh Annual Conference on Computational Learning Theory (New York, N.Y.), ACM, 1994, p. 299.

[9] P. Bartlett and M. Vidyasagar (eds.), *Learning theory special issue of Systems and Control Letters*, vol. 34, 1998.

[10] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long, *Characterizations of learnability for classes of* $\{0, \cdots, n\}$*-valued functions*, J. Comput. System Sci. **50** (1995), no. 1, 74–86.

[11] C. Bishop (ed.), *Neural networks and machine learning*, Computer and Systems Sciences, vol. 168, Springer Verlag, 1998.

[12] A. Blum and T. Mitchell, *Combining labeled and unlabeled data with co-training*, Proceedings of the 11th Annual Conference on Computational Learning Theory, 1998, pp. 92–100.

[13] M. Campi and P. Kumar, *Leaning dynamical systems in a stationary environment*, Systems and Control Letters **34** (1998), no. 3, 125–132.

[14] B. DasGupta and E. D. Sontag, *Sample complexity for learning recurrent perceptron mappings*, IEEE Trans. Inform. Theory **42** (1996), no. 5, 1479–1487.

[15] C. Dellacherie and P.-A. Meyer, *Probabilities and potential*, North-Holland, 1978.

[16] A. Dembo and O. Zeitouni, *Large deviations techniques*, Jones and Bartlett Publishers, 1993.

[17] L. Devroye, *Bounds for the uniform deviation of empirical measures*, Journal of Multivariate Analysis **12** (1982), 72–79.

[18] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*, corrected 2nd printing ed., Applications in Mathematics, Springer Verlag, 1997.

[19] P. Dupuis and R. S. Ellis, *A weak convergence approach to the theory of large deviations*, John Wiley & Sons, 1997.

[20] R. Ellis, *Entropy, large deviations, and statistical mechanics*, 1st ed., Comprehensive studies in mathematics, Springer Verlag, 1985.

[21] C.-N. Fiechter, *PAC adaptive control of linear systems*, Tenth Annual Conference on Computational Learning Theory, ACM, 1997.

[22] P. Goldberg and M. Jerrum, *Bounding the Vapnik-Chervonenkis dimension of concept classes parametrized by real numbers*, Machine Learning **18** (1995), 131–148.

[23] L. Gurvits, *A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces*, Proceedings of Algorithm Learning Theory, ALT-97, 1997, and as NECI Technical Report, 1997.

[24] W. Hoeffding, *Probability inequalities for sums of bounded random variables*, Journal of American Statistical Association **58** (1963), 13–30.

[25] T. Johansen and E. Weyer, *On convergence proofs in system identification – a general principle using ideas from learning theory*, Systems and Contol Letters **34** (1998), 85–92.

[26] P. Koiran and E. D. Sontag, *Vapnik-Chervonenkis dimension of recurrent neural networks*, Discrete Applied Mathematics **86** (1998), no. 1, 63–80.

[27] R. Koplon and E. D. Sontag, *Linear systems with sign-observations*, SIAM J. Control Optim. **31** (1993), 1245–1266.

[28] L. Ljung, *PAC-learning and symptotic system identification theory*, Proceedings of 35th IEEE Conf. on Decision and Control, Kobe, Japan, IEEE, 1996, pp. 2303–2307.

[29] R. Meir, *Performance bounds for nonlinear time series prediction*, Proc. Tenth Annual Conf. on Computational Learning Theory, 1997.

[30] D. Pollard, *Convergence of stochastic processes*, Springer series in statistics, Springer Verlag, 1984.

[31] J. Ratsaby and V. Maiorov, *On the value of partial information for learning from examples*, Journal of Complexity **13** (1997), no. 4, 509–544.

[32] N. Sauer, *On the density of families of sets*, Journal of Combinatorial Theory **13** (1972), 145–147.

[33] J Shawe-Taylor, M Anthony, and N. L. Biggs, *Bounding sample size with the Vapnik-Chervonenkis dimension*, Discrete Applied Mathematics **42** (1993), 65–73.

[34] J. Shawe-Taylor, P. Bartlett, R Williamson, and M. Anthony, *Structural risk minimization over data-dependent hierarchies*, NeuroCOLT Technical Report, NC-TR-96-053 (1996), available at `ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports` or `www.neurocolt.com`

[35] S. Shelah, *A combinatorial problem: Stability adn order for models and theories in infinitary languages*, Pacific Journal of Mathematics **41** (1972), 247–261.

[36] E. Sontag, *Feedforward nets for interpolation and classification*, J. Comp. Syst. Sci. **45** (1992), 20–48.

[37] V. Vapnik, *The nature of statistical learning theory*, Springer-Verlag, 1995.

[38] V Vapnik and A. Cervonenkis, *The uniform convergence of frequencies of the appearance of events to their probabilities*, Dokl. Akad. Nauk. SSSR (1968), in Russian.

[39] M. Vidyasagar, *A theory of learning and generalization; with applications to neural networks and control systems*, Springer-Verlag, 1997.

[40] ———, *Introduction to some statistical aspects of PAC learning theory*, Systems and Control Letters **34** (1998), no. 3, 115–124.

[41] R. Wenocur and R. Dudley, *Some special Vapnik-Chervonenkis classes*, Discrete Mathematics **33** (1981), 313–318.

[42] B. Yu, *Rates of convergence for empirical processes of stationary mixing sequences*, Ann. Probab. **22** (1994), no. 1, 94–116.

[43] A. Zador and B. Pearlmutter, *VC dimension of an integrate-and-fire neuron model*, Neural Computation **8** (1996), no. 3, 611–624.