Machine Learning Lecture Notes

Yevgeny Seldin

November 15, 2021

Foreword

The lecture notes are used as the primary teaching material for the following courses:

- Machine Learning A, Department of Computer Science, University of Copenhagen
- Machine Learning B, Department of Computer Science, University of Copenhagen.
- Machine Learning, Department of Computer Science, University of Copenhagen.
- Online and Reinforcement Learning, Department of Computer Science, University of Copenhagen.
- Advanced Topics in Machine Learning, Department of Computer Science, University of Copenhagen.

The notes are periodically updated (check the compilation date on the title page). The courses are co-taught by me, Christian Igel, Sadegh Talebi, and Fabian Gieseke. The notes only cover my part of the above courses.

I would like to thank all students who have pointed out typos and flaws in the lecture notes. There are certainly more and if you spot any, please, report them to me at seldin@di.ku.dk. Your feedback will serve future generations of students.

Contents

Chapter 1

Supervised Learning

The most basic and widespread form of machine learning is supervised learning. In the classical batch supervised learning setting the learner is given an annotated sample, which is used to derive a prediction rule for annotating new samples. We start with a simple informal example and then formalize the problem.

Let's say that we want to build a prediction rule that will use the average grade of a student in home assignments, say on a 100-points scale, to predict whether the student will pass the final exam. Such a prediction rule could be used for preliminary filtering of students to be allowed to take the final exam. The annotated sample could be a set of average grades of students from the previous year with indications of whether they have passed the final exam. The prediction rule could take a form of a threshold grade (a.k.a. decision stump), above which the student is expected to pass and below fail.

Now assume that we want to take a more refined approach and look into individual grades in each assignment, say, 5 assignments in total. For example, different assignments may have different relevance for the final exam or, maybe, some students may demonstrate progression throughout the course, which would mean that their early assignments should not be weighted equally with the later ones. In the refined approach each student can be represented by a point in a 5-dimensional space. The one-dimensional threshold could be replaced by a separating hyperplane, which separates the 5-dimensional space of grades into a linear subspace, where most students are likely to pass, and the complement, where they are likely to fail. An alternative approach is to look at "nearest neighbors" of a student in the space of grades. Given a grade profile of a student (the point representing the student in the 5-dimensional space) we look at students with the closest grade profile and see whether most of them passed or failed. This is known as the *K Nearest Neighbors* algorithm, where K is the number of neighbors we look at. But how many neighbors K should we look at? Considering the extremes gives some intuition about the problem. Taking just one nearest neighbor may be unreliable. For example, we could have a good student that accidentally failed the final exam and then all the neighbors will be marked as "expected to fail". Going in the other extreme and taking all the students in as neighbors is also undesirable, because effectively it will ignore the individual profile altogether. So a good value of K should be somewhere between 1 and n, where n is the size of the annotated set. But how to find it? Well, read on and you will learn how to approach this question formally.

1.1 The Supervised Learning Setting

We start with a bunch of notations and then illustrate them with examples.

- $\mathcal{X}$ - the sample space.
- $\mathcal{Y}$ - the label space.
- $X \in \mathcal{X}$ - unlabeled sample.
- $(X, Y) \in (\mathcal{X} \times \mathcal{Y})$ - labeled sample.
- $S = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ - a training set. We assume that (X_i, Y_i) pairs in S are sampled i.i.d. according to an unknown, but fixed distribution $p(X, Y)$.

- $h : \mathcal{X} \to \mathcal{Y}$ - a hypothesis, which is a function from $\mathcal{X}$ to $\mathcal{Y}$.
- $\mathcal{H}$ - a hypothesis set.
- $\ell(Y', Y)$ - the loss function for predicting Y' instead of Y.
- $\hat{L}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$ - the empirical loss (a.k.a. error or risk) of h on S. (In many textbooks S is omitted from the notation and $\hat{L}(h)$ or $\hat{L}_n(h)$ is used to denote $\hat{L}(h, S)$.)
- $L(h) = \mathbb{E}[\ell(h(X), Y)]$ - the expected loss (a.k.a. error or risk) of h, where the expectation is taken with respect to $p(X, Y)$.

The Learning Protocol

The classical supervised learning acts according to the following protocol:

1. The learner gets a training set S of size n sampled i.i.d. according to $p(X, Y)$.
2. The learner returns a prediction rule h.
3. New instances (X, Y) are sampled according to $p(X, Y)$, but only X is observed and h is used to predict the unobserved Y.

The goal of the learner is to return h that minimizes $L(h)$, which is the expected error on new samples.

Examples - Sample and Label Spaces

Let's say that we want to predict person's height based on age, gender, and weight. Then $\mathcal{X} = \mathbb{N} \times \{\pm 1\} \times \mathbb{R}$ and $\mathcal{Y} = \mathbb{R}$. If we want to predict gender based on age, weight, and height, then $\mathcal{X} = \mathbb{N} \times \mathbb{R} \times \mathbb{R}$ and $\mathcal{Y} = \{\pm 1\}$. If we want to predict the height of a baby at the age of 4 years based on his or her height at the ages of 1, 2, and 3 years, then $\mathcal{X} = \mathbb{R}^3$ and $\mathcal{Y} = \mathbb{R}$.

1.1.1 Classification, Regression, and Other Supervised Learning Problems

The most widespread forms of supervised learning are classification and regression. We also mention a few more, mainly to show that the supervised learning setting is much richer.

Classification A supervised learning problem is a classification problem when the output (label) space $\mathcal{Y}$ is binary. The goal of the learning algorithm is to separate between two classes: yes or no; good or bad; healthy or sick; male or female; etc. Most often the translation of the binary label into numerical representation is done by either taking $\mathcal{Y} = \{\pm 1\}$ or $\mathcal{Y} = \{0, 1\}$. Sometimes the setting is called *binary classification* to emphasize that $\mathcal{Y}$ takes just two values.

Regression A supervised learning problem is a regression problem when the output space $\mathcal{Y} = \mathbb{R}$. For example, prediction of person's height would be a regression problem.

Multiclass Classification When $\mathcal{Y}$ consists of a finite and typically unordered and relatively small set of values, the corresponding supervised learning problem is called multiclass classification. For example, prediction of a study program a student will apply for based on his or her grades would be a multiclass classification problem. Finite ordered output spaces, for example, prediction of age or age group can also be modeled as multiclass classification, but it may be possible to exploit the structure of $\mathcal{Y}$ to obtain better solutions. For example, it may be possible to exploit the fact that ages 22 and 23 are close together, whereas 22 and 70 are far apart; therefore, it may be possible to share some information between close ages, as well as exploit the fact that predicting 22 instead of 23 is not such a big mistake as predicting 22 instead of 70. Depending on the setting, it may be preferable to model prediction of ordered sets as regression rather than multiclass classification.

Structured Prediction Consider the problem of machine translation. An algorithm gets a sentence in English as an input and should produce a sentence in Danish as an output. In this case the output (the sentence in Danish) is not merely a number, but a structured object and such prediction problems are known as structured prediction.

1.1.2 The Loss Function $\ell(Y', Y)$

The loss function (a.k.a. the error function) encodes how much the user of an algorithm cares about various kinds of mistakes. Most literature on binary classification, including these lecture notes, uses the *zero-one loss* defined by

$$\ell(Y', Y) = \mathbb{1}(Y' \neq Y) = \begin{cases} 1, & \text{if } Y' \neq Y \\ 0, & \text{otherwise,} \end{cases}$$

where $\mathbb{1}$ is the indicator function. Common loss functions in regression are the *square loss*

$$\ell(Y', Y) = (Y' - Y)^2$$

and the *absolute loss*

$$\ell(Y', Y) = |Y' - Y|.$$

The above loss functions are convenient general choices, but not necessarily the right choice for a particular application. For example, imagine that you design an algorithm for fire alarm that predicts "fire / no fire". Assume that the cost of a house is 3,000,000 DKK and the cost of calling in a fire brigade is 2,000 DKK. Then the loss function would be

$\ell(Y', Y) =$

Y' \ Y	no fire	fire
no fire	0	3,000,000
fire	2,000	0

The loss for making the correct prediction is zero, but the loss of *false positive* (predicting fire when in reality there is no fire) and *false negative* (predicting no fire when the reality is fire) are not symmetric anymore.

Put attention that the loss depends on how the predictions are used and the loss table depends on the user. For example, if the same alarm is installed in a house that is worth 10,000,000 DKK, the ratio between the cost of false positives and false negatives will be very different and, as a result, the optimal prediction strategy will not necessarily be the same.

1.2 K Nearest Neighbors for Binary Classification

One of the simplest algorithms for binary classification is K Nearest Neighbors (K-NN). The algorithm is based on an externally provided distance function $d(\mathbf{x}, \mathbf{x}')$ that computes distances between pairs of points $\mathbf{x}$ and $\mathbf{x}'$. For example, for points in $\mathbb{R}^d$ the distance could be the Euclidean distance $d(\mathbf{x}, \mathbf{x}') = \|\mathbf{x}-\mathbf{x}'\| = \sqrt{\sum_{j=1}^{d}(x_j - x'_j)^2} = \sqrt{(\mathbf{x}-\mathbf{x}')^T(\mathbf{x}-\mathbf{x}')}$, where $\mathbf{x} = (x_1, \ldots, x_d)$ and x_j is the j-th coordinate of vector $\mathbf{x}$. Other choices of distance measures are possible and, in general, lead to different predictions. The choice of the distance measure $d(\mathbf{x}, \mathbf{x}')$ is the key for success or failure of K-NN, but we leave the topic of selection of d outside the scope of the lecture notes.

K-NN algorithm takes as input a set of training points $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and predicts the label of a target point $\mathbf{x}$ based on the majority vote of K points from S, which are the closest to $\mathbf{x}$ in terms of the distance measure $d(\mathbf{x}_i, \mathbf{x})$.

The ordering of d_i-s in Step 3 is identical to the ordering of d_i^2 and for the Euclidean distance we can save the computation of the square root by working with squared distances.

The hypothesis space $\mathcal{H}$ is implicit in the K-NN algorithm. It is the space of all possible partitions of the sample space $\mathcal{X}$. The output hypothesis h is parametrized by all training points $h_S = h_{\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}}$. In the sequel we will see other prediction rules that operate with more explicit hypothesis spaces, for example, a space of all linear separators.

Algorithm 1 K Nearest Neighbors (K-NN) for Binary Classification with $\mathcal{Y} = \{\pm 1\}$

1: **Input:** A set of labeled points $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ and a target point $\mathbf{x}$ that has to be classified.
2: Calculate the distances $d_i = d(\mathbf{x}_i, \mathbf{x})$.
3: Sort d_i-s in ascending order and let $\sigma : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be the corresponding permutation of indices. In other words, for any pair of indices $i < j$ we should have $d_{\sigma(i)} \leq d_{\sigma(j)}$.
4: The output of K-NN is $y = \text{sign}\left(\sum_{i=1}^{K} y_{\sigma(i)}\right)$. It is the majority vote of K points that are the closest to $\mathbf{x}$. Note that we can calculate the output of K-NN for all K in one shot.

1.2.1 How to Pick K in K-NN?

One of the key questions in K-NN is how to pick K. It is instructive to consider the extreme cases to gain some intuition. In 1-NN the prediction is based on a single sample $(\mathbf{x}_i, y_i)$ which happens to be closest to the target point $\mathbf{x}$. This may not be the best thing to do. Imagine that you are admitted to a hospital and a diagnostic system determines whether you are healthy or sick based on a single annotated patient that has the symptoms closest to yours (in distance measure d). You would likely prefer to be diagnosed based on the majority of diagnoses of several patients with similar symptoms. At the other extreme, in n-NN, where n is the number of samples in S, the prediction is based on the majority of labels y_i within the sample S, without even taking any particular $\mathbf{x}$ into account. So the desirable K is somewhere between 1 and n, but how to find it?

Let $h_{\texttt{K-NN}}$ denote the prediction rule of K-NN. As K goes from 1 to n, K-NN provides n different prediction rules, $h_{\texttt{1-NN}}$, $h_{\texttt{2-NN}}$, $\ldots$, $h_{\texttt{n-NN}}$ (or half of that if we only take the odd values of K). Recall that we are interested in finding K that minimizes the expected loss $L(h_{\texttt{K-NN}})$ and that $L(h_{\texttt{K-NN}})$ is unobserved. We can calculate the empirical loss $\hat{L}(h_{\texttt{K-NN}}, S)$ for any K. However, $\hat{L}(h_{\texttt{1-NN}}, S)$ is always zero[1] and in general the empirical error of K-NN is an underestimate of its expected error and we need other tools to estimate $L(h_{\texttt{K-NN}})$. We start developing these tools in the next section and continue throughout the lecture notes.

1.3 Validation

Whenever we select a hypothesis $\hat{h}^*_S$ out of a hypothesis set $\mathcal{H}$ based on empirical performances $\hat{L}(h, S)$, the empirical performance $\hat{L}(\hat{h}^*_S, S)$ becomes a biased estimate of $L(\hat{h}^*_S)$. This is clearly observed in 1-NN, where $\hat{L}(h_{\texttt{1-NN}}, S) = 0$, but $L(h_{\texttt{1-NN}})$ is most often not zero (we remind that the hypothesis space in 1-NN is the space of all possible partitions of the sample space $\mathcal{X}$ and $h_{\texttt{1-NN}}$ is the hypothesis that achieves the minimal empirical error in this space). The reason is that when we do the selection we pick $\hat{h}^*_S$ that is best suited for S (it achieves the minimal $\hat{L}(h, S)$ out of all h in $\mathcal{H}$). Therefore, from the perspective of $\hat{h}^*_S$ the new samples (X, Y) are not "similar" to the samples (X_i, Y_i) in S. A bit more precisely, (X, Y) is not exchangeable with (X_i, Y_i), because if we would exchange (X_i, Y_i) with (X, Y) it is likely that $\hat{h}^*_S$, the hypothesis that minimizes $\hat{L}(h, S)$, would be different. Again, this is very clear in 1-NN: if we change one sample (X_i, Y_i) in S we get a different prediction rule $h_{\texttt{1-NN}}$. We get back to this topic in much more details in Chapter 3 after we develop some mathematical tools for analyzing the bias in Chapter 2. For now we present a simple solution for estimating $L(\hat{h}^*_S)$ and motivate why we need the tools from Chapter 2.

The solution is to split the sample set S into training set $S_{\texttt{train}}$ and validation set $S_{\texttt{val}}$. We can then find the best hypothesis for the training set, $h^*_{S_{\texttt{train}}}$, and validate it on the validation set by computing $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$. Note that from the perspective of $h^*_{S_{\texttt{train}}}$ the samples in $S_{\texttt{val}}$ are exchangeable with any new samples (X, Y). If we exchange $(X_i, Y_i) \in S_{\texttt{val}}$ with another sample (X, Y) coming from the same distribution, $h^*_{S_{\texttt{train}}}$ will stay the same and in expectation $\mathbb{E}\left[\ell(h^*_{S_{\texttt{train}}}(X_i), Y_i)\right] = \mathbb{E}\left[\ell(h^*_{S_{\texttt{train}}}(X), Y)\right]$, meaning that on average $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ will also stay the same (only on average, the exact value may change). Therefore, $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ is an unbiased estimate of $L(h^*_{S_{\texttt{train}}})$. (We get back to this point in much more details in Chapter 3.)

[1] This is because the closest point in S to a sample point $\mathbf{x}_i$ is $\mathbf{x}_i$ itself and we assume that S includes no identical points with dissimilar labels, which is a reasonable assumption if $\mathcal{X} = \mathbb{R}^d$.

Now we get to the question of how to split S into $S_{\texttt{train}}$ and $S_{\texttt{val}}$, and again it is very instructive to consider the extreme cases. Imagine that we keep a single sample for validation and use the remaining $n-1$ samples for training. Let's say that we keep the last sample, (X_n, Y_n), for validation, then $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}}) = \ell(h^*_{S_{\texttt{train}}}(X_n), Y_n)$ and in the case of zero-one loss it is either zero or one. Even though $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ is an unbiased estimate of $L(h^*_{S_{\texttt{train}}})$, it clearly does not represent it well. At the other extreme, if we keep $n-1$ points for validation and use the single remaining point for training we run into a different kind of problem: a classifier trained on a single point is going to be extremely weak. Let's say that we have used the first point, (X_1, Y_1), for training. In the case of K-NN classifier, as well as most other classifiers, $h^*_{S_{\texttt{train}}}$ will always predict Y_1, no matter what input it gets. The validation error $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ will be a very good estimate of $L(h^*_{S_{\texttt{train}}})$, but this is definitely not a classifier we want.

So how many samples from S should go into $S_{\texttt{train}}$ and how many into $S_{\texttt{val}}$? Currently there is no "gold answer" to this question, but in Chapters 2 and 3 we develop mathematical tools for intelligent reasoning about it. An important observation to make is that for h independent of (X, Y) the zero-one loss $\ell(h(X), Y)$ is a Bernoulli random variable with bias $\mathbb{P}(\ell(h(X), Y) = 1) = L(h)$. Furthermore, when h is independent of a set of samples $\{(X_1, Y_1), \ldots, (X_m, Y_m)\}$ (i.e., these samples are not used for selecting h), the losses $\ell(h(X_i), Y_i)$ are independent identically distributed (i.i.d.) Bernoulli random variables with bias $L(h)$. Therefore, when $S_{\texttt{val}}$ is of size m, the validation loss $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ is an average of m i.i.d. Bernoulli random variables with bias $L(h^*_{S_{\texttt{train}}})$. The validation loss $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ is observed, but the expected loss that we are actually interested in is unobserved. One of the key questions that we are interested in is how far $\hat{L}(h^*_{S_{\texttt{train}}}, S_{\texttt{val}})$ can be from $L(h^*_{S_{\texttt{train}}})$. We have already seen that $m = 1$ is too little. But how large should it be, 10, 100, 1000? Essentially this question is equivalent to asking how many times do we need to flip a biased coin in order to get a satisfactory estimate of its bias. In Chapter 2 we develop concentration of measure inequalities that answer this question.

Another technical question is which samples should go into $S_{\texttt{train}}$ and which into $S_{\texttt{val}}$? From the theoretical perspective we assume that S is sampled i.i.d. and, therefore, it does not matter. We can take the first $n-m$ samples into $S_{\texttt{train}}$ and the last m into $S_{\texttt{val}}$ or split in any other way. From a practical perspective the samples may actually not be i.i.d. and there could be some parameter that has influenced their order in S. For example, they could have been ordered alphabetically. Therefore, from a practical perspective it is desirable to take a random permutation of S before splitting, unless the order carries some information we would like to preserve. For example, if S is a time-ordered series of product reviews and we would like to build a classifier that classifies them into positive and negative, we may want to get an estimate of temporal variation and keep the order when we do the split, i.e., train on the earlier samples and validate on the later.

1.3.1 Test Set: It's not about how you call it, it's about how you use it!

Assume that we have split S into $S_{\texttt{train}}$ and $S_{\texttt{val}}$; we have trained $h_{\texttt{1-NN}}, \ldots, h_{\texttt{n-NN}}$ on $S_{\texttt{train}}$; we calculated $\hat{L}(h_{\texttt{1-NN}}, S_{\texttt{val}}), \ldots, \hat{L}(h_{\texttt{n-NN}}, S_{\texttt{val}})$ and picked the value K^* that minimizes $\hat{L}(h_{\texttt{K-NN}}, S_{\texttt{val}})$. Is $\hat{L}(h_{\texttt{K*-NN}}, S_{\texttt{val}})$ an unbiased estimate of $L(h_{\texttt{K*-NN}})$?

This is probably one of the most conceptually difficult points about validation, at least when you encounter it for the first time. While for each $h_{\texttt{K-NN}}$ individually $\hat{L}(h_{\texttt{K-NN}}, S_{\texttt{val}})$ is an unbiased estimate of $L(h_{\texttt{K-NN}})$, the validation loss $\hat{L}(h_{\texttt{K*-NN}}, S_{\texttt{val}})$ is a *biased* estimate of $L(h_{\texttt{K*-NN}})$. This is because $S_{\texttt{val}}$ was used for selection of K^* and, therefore, $h_{\texttt{K*-NN}}$ depends on $S_{\texttt{val}}$. So if we want to get an unbiased estimate of $L(h_{\texttt{K*-NN}})$ we have to reserve some "fresh" data for that. So we need to split S into $S_{\texttt{train}}$, $S_{\texttt{val}}$, and $S_{\texttt{test}}$; train the K-NN classifiers on $S_{\texttt{train}}$; pick the best K^* based on $\hat{L}(h_{\texttt{K-NN}}, S_{\texttt{val}})$; and then compute $\hat{L}(h_{\texttt{K*-NN}}, S_{\texttt{test}})$ to get an unbiased estimate of $L(h_{\texttt{K*-NN}})$.

It's not about how you call it, it's about how you use it! Some people think that if you call some data a test set it automatically makes loss estimates on this set unbiased. This is not true. Imagine that you have split S into $S_{\texttt{train}}$, $S_{\texttt{val}}$, and $S_{\texttt{test}}$; you trained K-NN on $S_{\texttt{train}}$, picked the best value K^* using $S_{\texttt{val}}$, and estimated the loss of $h_{\texttt{K*-NN}}$ on $S_{\texttt{test}}$. And now you are unhappy with the result and you want to try a different learning method, say a neural network. You go through the same steps: you train networks with various parameter settings on $S_{\texttt{train}}$, you validate them on $S_{\texttt{val}}$, and you pick the best parameter set θ^* based on the validation loss. Finally, you compute the test loss of the neural network parametrized by θ^* on $S_{\texttt{test}}$. It happens to be lower than the test loss of K^*-NN and you decide to go

with the neural network. Does the empirical loss of the neural network on $S_{\texttt{test}}$ represent an unbiased estimate of its expected loss? No! Why? Because our choice to pick the neural network was based on its superior performance relative to $h_{\texttt{K*-NN}}$ on $S_{\texttt{test}}$, so $S_{\texttt{test}}$ was used in selection of the neural network. Therefore, there is dependence between $S_{\texttt{test}}$ and the hypothesis we have selected, and the loss on $S_{\texttt{test}}$ is biased. If we want to get an unbiased estimate of the loss we have to find new "fresh" data or reserve such data from the start and keep it in a locker until the final evaluation moment. Alternatively, we can correct for the bias and in Chapter 3 we will learn some tools for making the correction. The main take-home message is: ***It is not about how you call a data set, S_{train}, S_{val}, or S_{test}, it is the way you use it which determines whether you get unbiased estimates or not!*** In some cases it is possible to get unbiased estimates or to correct for the bias already with $S_{\texttt{train}}$, and sometimes there is bias even on $S_{\texttt{test}}$ and we need to correct for that.

1.3.2 Cross-Validation

Sometimes it feels wasteful to use only part of the data for training and part for validation. A *heuristic* way around it is cross-validation. In the standard N-fold cross-validation setup the data S are split into N non-overlapping folds $S_1, \ldots, S_N$. Then for $i \in \{1, \ldots, N\}$ we train on all folds except the i-th and validate on S_i. We then take the average of the N validation errors and pick the parameter that achieves the minimum (for example, the best K in K-NN). Finally, we train a model with the best parameter we have selected in the cross-validation procedure (for example the best K^* in K-NN) using all the data S.

The standard cross-validation procedure described above is a heuristic and has no theoretical guarantees. It is fairly robust and widely used in practice, but it is possible to construct examples, where it fails. In Chapter 3 we describe a modification of the cross-validation procedure, which comes with theoretical generalization guarantees and is empirically competitive with the standard cross-validation procedure.

1.4 Perceptron - Basic Algorithm for Linear Classification

Linear classification is another basic family of classification strategies. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{\pm 1\}$. A hyperplane in $\mathbb{R}^d$ is described by a tuple $(\mathbf{w}, b)$, where $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$. The points $\mathbf{x}$ on the hyperplane are described by the equation

$$\mathbf{w}^T\mathbf{x} + b = 0.$$

A linear classifier $h = (\mathbf{w}, b)$ assigns label $+1$ to all points on the "positive" side of the hyperplane and -1 on the "negative" side of the hyperplane. Specifically,

$$h(\mathbf{x}) = \text{sign}\left(\mathbf{w}^T\mathbf{x} + b\right).$$

Homogeneous classifiers We distinguish between *homogeneous* linear classifiers and non-homogeneous linear classifiers. A homogeneous linear classifier is described by a hyperplane passing through the origin. From the mathematical point of view it means that $b = 0$.

We note that any linear classifier in $\mathbb{R}^d$ can be transformed into a homogeneous linear classifier in $\mathbb{R}^{d+1}$ by the following transformation

$$\mathbf{x} \to (\mathbf{x}; 1)$$
$$\{\mathbf{w}, b\} \to (\mathbf{w}; b)$$

(where by ";" we mean that we append a row to a column vector). In other words, we append "1" to the $\mathbf{x}$ vector and combine $\mathbf{w}$ and b into one vector in $\mathbb{R}^{d+1}$. Note that $\mathbf{w}^T\mathbf{x} + b = (\mathbf{w}; b)^T(\mathbf{x}; 1)$ and, therefore, the predictions of the transformed model are identical to predictions of the original model. Through this transformation any learning algorithm for homogeneous classifiers can be directly applied to learning non-homogeneous classifiers.

Hypothesis space The hypothesis space in linear classification is the space of all possible separating hyperplanes. If we are talking about homogeneous linear classifiers then it is restricted to hyperplanes passing through the origin. Thus, for homogeneous linear classifiers $\mathcal{H} = \mathbb{R}^d$ and for general linear classifiers $\mathcal{H} = \mathbb{R}^{d+1}$.

Perceptron algorithm Perceptron is the simplest algorithm for learning homogeneous separating hyperplanes. It operates under the ***assumption that the data are separable by a homogeneous hyperplane***, meaning that there exists a hyperplane passing through the origin that perfectly separates positive points from negative.

Algorithm 2 Perceptron

1: **Input:** A training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
2: **Initialization:** $\mathbf{w}_1 = \mathbf{0}$ (where $\mathbf{0}$ is the zero vector)
3: $t = 1$
4: **while** exists $(\mathbf{x}_{i_t}, y_{i_t})$, such that $y_{i_t}(\mathbf{w}_t^T \mathbf{x}_{i_t}) \leq 0$ **do**
5: $\quad \mathbf{w}_{t+1} = \mathbf{w}_t + y_{i_t}\mathbf{x}_{i_t}$
6: $\quad t = t + 1$
7: **end while**
8: **Return:** $\mathbf{w}_t$

Note that a point $(\mathbf{x}, y)$ is classified correctly if $y\mathbf{w}^T\mathbf{x} > 0$ and misclassified if $y\mathbf{w}^T\mathbf{x} \leq 0$. Thus, the selection step (line 4 in the pseudocode) picks a misclassified point, as long as there exists such. The update step (line 5 in the pseudocode) rotates the hyperplane $\mathbf{w}$, so that the classification is "improved". Specifically, the following property is satisfied: if $(\mathbf{x}_{i_t}, y_{i_t})$ is the point selected at step t then $y_{i_t}\mathbf{w}_{t+1}^T\mathbf{x}_{i_t} > y_{i_t}\mathbf{w}_t^T\mathbf{x}_{i_t}$ (verification of this property is left as an exercise to the reader). Note this property does not guarantee that after the update $\mathbf{w}_{t+1}$ will classify $(\mathbf{x}_{i_t}, y_{i_t})$ correctly. But it will rotate in the right direction and after sufficiently many updates $(\mathbf{x}_{i_t}, y_{i_t})$ will end up on the right side of the hyperplane. Also note that while the classification of $(\mathbf{x}_{i_t}, y_{i_t})$ is improved, it may go the opposite way for other points. As long as the data are linearly separable, the algorithm will eventually find the separation.

The algorithm does not specify the order in which misclassified points are selected. Two natural choices are sequential and random. We leave it as an exercise to the reader to check which of the two choices leads to faster convergence of the algorithm.

Chapter 2

Concentration of Measure Inequalities

Concentration of measure inequalities are one of the main tools for analyzing learning algorithms. This chapter is devoted to a number of concentration of measure inequalities that form the basis for the results discussed in later chapters.

2.1 Markov's Inequality

Markov's Inequality is the simplest and relatively weak concentration inequality. Nevertheless, it forms the basis for many much stronger inequalities that we will see in the sequel.

Theorem 2.1 (Markov's Inequality)**.** *For any non-negative random variable X and $\varepsilon > 0$:*

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}.$$

Proof. Define a random variable $Y = \mathbb{1}(X \geq \varepsilon)$ to be the indicator function of whether X exceeds ε. Then $Y \leq \frac{X}{\varepsilon}$ (see Figure 2.1). Since Y is a Bernoulli random variable, $\mathbb{E}[Y] = \mathbb{P}(Y = 1)$ (see Appendix B). We have:

$$\mathbb{P}(X \geq \varepsilon) = \mathbb{P}(Y = 1) = \mathbb{E}[Y] \leq \mathbb{E}\left[\frac{X}{\varepsilon}\right] = \frac{\mathbb{E}[X]}{\varepsilon}.$$

Check yourself: where in the proof do we use non-negativity of X and strict positiveness of ε? □

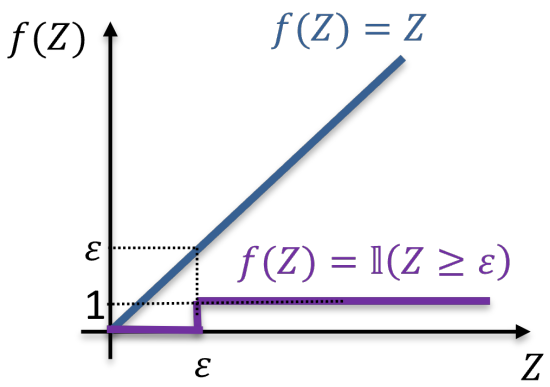


Figure 2.1: Relation between the identity function and the indicator function.

By denoting the right hand side of Markov's inequality by δ we obtain the following equivalent statement. For any non-negative random variable X:

$$\mathbb{P}\left(X \geq \frac{1}{\delta}\mathbb{E}\left[X\right]\right) \leq \delta.$$

Example. We would like to bound the probability that we flip a fair coin 10 times and obtain 8 or more heads. Let $X_1, \ldots, X_{10}$ be i.i.d. Bernoulli random variables with bias $\frac{1}{2}$. The question is equivalent to asking what is the probability that $\sum_{i=1}^{10} X_i \geq 8$. We have $\mathbb{E}\left[\sum_{i=1}^{10} X_i\right] = 5$ (the reader is invited to prove this statement formally) and by Markov's inequality

$$\mathbb{P}\left(\sum_{i=1}^{10} X_i \geq 8\right) \leq \frac{\mathbb{E}\left[\sum_{i=1}^{10} X_i\right]}{8} = \frac{5}{8}.$$

We note that even though Markov's inequality is weak, there are situations in which it is tight. We invite the reader to construct an example of a random variable for which Markov's inequality is tight.

2.2 Chebyshev's Inequality

Our next stop is Chebyshev's inequality, which exploits variance to obtain tighter concentration.

Theorem 2.2 (Chebyshev's inequality)**.** *For any $\varepsilon > 0$*

$$\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) \leq \frac{\mathrm{Var}\left[X\right]}{\varepsilon^2}.$$

Proof. The proof uses a transformation of a random variable. We have that $\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) = \mathbb{P}\Big((X - \mathbb{E}\left[X\right])^2 \geq \varepsilon^2\Big)$, because the first statement holds if and only if the second holds. In addition, using Markov's inequality and the fact that $(X - \mathbb{E}\left[X\right])^2$ is a non-negative random variable we have

$$\mathbb{P}(|X - \mathbb{E}\left[X\right]| \geq \varepsilon) = \mathbb{P}\Big((X - \mathbb{E}\left[X\right])^2 \geq \varepsilon^2\Big) \leq \frac{\mathbb{E}\left[(X - \mathbb{E}\left[X\right])^2\right]}{\varepsilon^2} = \frac{\mathrm{Var}\left[X\right]}{\varepsilon^2}.$$

Check yourself: where in the proof did we use the positiveness of ε? □

In order to illustrate the relative advantage of Chebyshev's inequality compared to Markov's consider the following example. Let $X_1, \ldots, X_n$ be n independent identically distributed Bernoulli random variables and let $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ be their average. We would like to bound the probability that $\hat{\mu}_n$ deviates from $\mathbb{E}\left[\hat{\mu}_n\right]$ by more than ε (this is the central question in machine learning). We have $\mathbb{E}\left[\hat{\mu}_n\right] = \mathbb{E}\left[X_1\right] = \mu$ and by independence of X_i-s and Theorem B.26 we have $\mathrm{Var}\left[\hat{\mu}_n\right] = \frac{1}{n^2}\mathrm{Var}\left[n\hat{\mu}_n\right] = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}\left[X_i\right] = \frac{1}{n}\mathrm{Var}\left[X_1\right]$. By Markov's inequality

$$\mathbb{P}(\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right] \geq \varepsilon) = \mathbb{P}(\hat{\mu}_n \geq \mathbb{E}\left[\hat{\mu}_n\right] + \varepsilon) \leq \frac{\mathbb{E}\left[\hat{\mu}_n\right]}{\mathbb{E}\left[\hat{\mu}_n\right] + \varepsilon} = \frac{\mathbb{E}\left[X_1\right]}{\mathbb{E}\left[X_1\right] + \varepsilon}.$$

Note that as n grows the inequality stays the same. By Chebyshev's inequality we have

$$\mathbb{P}(\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right] \geq \varepsilon) \leq \mathbb{P}(|\hat{\mu}_n - \mathbb{E}\left[\hat{\mu}_n\right]| \geq \varepsilon) \leq \frac{\mathrm{Var}\left[\hat{\mu}_n\right]}{\varepsilon^2} = \frac{\mathrm{Var}\left[X_1\right]}{n\varepsilon^2}.$$

Note that as n grows the right hand side of the inequality decreases at the rate of $\frac{1}{n}$. Thus, in this case Chebyshev's inequality is much tighter than Markov's and it illustrates that as the number of random variables grows the probability that their average significantly deviates from the expectation decreases. In the next section we show that this probability actually decreases at an exponential rate.

2.3 Hoeffding's Inequality

Hoeffding's inequality is a much more powerful concentration result.

Theorem 2.3 (Hoeffding's Inequality)**.** *Let $X_1, \ldots, X_n$ be independent real-valued random variables, such that for each $i \in \{1, \ldots, n\}$ there exist $a_i \leq b_i$, such that $X_i \in [a_i, b_i]$. Then for every $\varepsilon > 0$:*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2} \tag{2.1}$$

and

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \leq -\varepsilon\right) \leq e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.2}$$

By taking a union bound of the events in (2.1) and (2.2) we obtain the following corollary.

Corollary 2.4. *Under the assumptions of Theorem 2.3:*

$$\mathbb{P}\left(\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| \geq \varepsilon\right) \leq 2e^{-2\varepsilon^2 / \sum_{i=1}^{n}(b_i - a_i)^2}. \tag{2.3}$$

Equations (2.1) and (2.2) are known as "one-sided Hoeffding's inequalities" and (2.3) is known as "two-sided Hoeffding's inequality".

If we assume that X_i-s are identically distributed and belong to the $[0, 1]$ interval we obtain the following corollary.

Corollary 2.5. *Let $X_1, \ldots, X_n$ be independent random variables, such that $X_i \in [0, 1]$ and $\mathbb{E}[X_i] = \mu$ for all i, then for every $\varepsilon > 0$:*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2} \tag{2.4}$$

and

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}. \tag{2.5}$$

Recall that by Chebyshev's inequality $\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges to μ at the rate of n^{-1}. Hoeffding's inequality demonstrates that the convergence is actually much faster, at least at the rate of e^{-n}.

The proof of Hoeffding's inequality is based on Hoeffding's lemma.

Lemma 2.6 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $X \in [a, b]$. Then for any $\lambda \in \mathbb{R}$:*

$$\mathbb{E}\left[e^{\lambda X}\right] \leq e^{\lambda \mathbb{E}[X] + \frac{\lambda^2 (b-a)^2}{8}}.$$

The function $f(\lambda) = \mathbb{E}\left[e^{\lambda X}\right]$ is known as the *moment generating function* of X, since $f'(0) = \mathbb{E}[X]$, $f''(0) = \mathbb{E}\left[X^2\right]$, and, more generally, $f^{(k)}(0) = \mathbb{E}\left[X^k\right]$. We provide the proof of the lemma immediately after the proof of Theorem 2.3.

Proof of Theorem 2.3. We prove the first inequality in Theorem 2.3. The second inequality follows by applying the first inequality to $-X_1, \ldots, -X_n$. The proof is based on Chernoff's bounding technique. For any $\lambda > 0$ the following holds:

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \geq \varepsilon\right) = \mathbb{P}\left(e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)} \geq e^{\lambda\varepsilon}\right) \leq \frac{\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)}\right]}{e^{\lambda\varepsilon}},$$

where the first step holds since $e^{\lambda x}$ is a monotonously increasing function for $\lambda > 0$ and the second step holds by Markov's inequality. We now take a closer look at the nominator:

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right]\right)}\right] &= \mathbb{E}\left[e^{\left(\sum_{i=1}^n \lambda(X_i - \mathbb{E}[X_i])\right)}\right] \\
&= \mathbb{E}\left[\prod_{i=1}^n e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \\
&= \prod_{i=1}^n \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \qquad (2.6) \\
&\leq \prod_{i=1}^n e^{\lambda^2 (b_i - a_i)^2/8} \qquad (2.7) \\
&= e^{(\lambda^2/8)\sum_{i=1}^n (b_i - a_i)^2},
\end{aligned}$$

where (2.6) holds since $X_1, \ldots, X_n$ are independent and (2.7) holds by Hoeffding's lemma applied to a random variable $Z_i = X_i - \mathbb{E}[X_i]$ (note that $\mathbb{E}[Z_i] = 0$ and that $Z_i \in [a_i - \mu_i, b_i - \mu_i]$ for $\mu_i = \mathbb{E}[X_i]$). *Put attention to the crucial role that independence of $X_1, \ldots, X_n$ plays in the proof! Without independence we would not have been able to exchange the expectation with the product and the proof would break down!* To complete the proof we substitute the bound on the expectation into the previous calculation and obtain:

$$\mathbb{P}\left(\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] \geq \varepsilon\right) \leq e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon}.$$

This expression is minimized by

$$\lambda^* = \arg\min_\lambda e^{(\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon} = \arg\min_\lambda \left((\lambda^2/8)\left(\sum_{i=1}^n (b_i - a_i)^2\right) - \lambda\varepsilon\right) = \frac{4\varepsilon}{\sum_{i=1}^n (b_i - a_i)^2}.$$

It is important to note that the best choice of λ does not depend on the sample. In particular, it allows to fix λ before observing the sample. By substituting λ^* into the calculation we obtain the result of the theorem. □

Proof of Lemma 2.6. Note that

$$\mathbb{E}\left[e^{\lambda X}\right] = \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X]) + \lambda\mathbb{E}[X]}\right] = e^{\lambda\mathbb{E}[X]} \times \mathbb{E}\left[e^{\lambda(X - \mathbb{E}[X])}\right].$$

Hence, it is sufficient to show that for any random variable Z with $\mathbb{E}[Z] = 0$ and $Z \in [a, b]$ we have:

$$\mathbb{E}\left[e^{\lambda Z}\right] \leq e^{\lambda^2 (b-a)^2/8}.$$

By convexity of the exponential function, for $z \in [a, b]$ we have:

$$e^{\lambda z} \leq \frac{z - a}{b - a} e^{\lambda b} + \frac{b - z}{b - a} e^{\lambda a}.$$

Let $p = -a/(b - a)$. Then:

$$\begin{aligned}
\mathbb{E}\left[e^{\lambda Z}\right] &\leq \mathbb{E}\left[\frac{Z - a}{b - a} e^{\lambda b} + \frac{b - Z}{b - a} e^{\lambda a}\right] \\
&= \frac{\mathbb{E}[Z] - a}{b - a} e^{\lambda b} + \frac{b - \mathbb{E}[Z]}{b - a} e^{\lambda a} \\
&= \frac{-a}{b - a} e^{\lambda b} + \frac{b}{b - a} e^{\lambda a} \\
&= \left(1 - p + p e^{\lambda(b-a)}\right) e^{-p\lambda(b-a)} \\
&= e^{\phi(u)},
\end{aligned}$$

where $u = \lambda(b-a)$ and $\phi(u) = -pu + \ln\left(1 - p + pe^u\right)$ and we used the fact that $\mathbb{E}[Z] = 0$. It is easy to verify that the derivative of ϕ is

$$\phi'(u) = -p + \frac{p}{p + (1-p)e^{-u}}$$

and, therefore, $\phi(0) = \phi'(0) = 0$. Furthermore,

$$\phi''(u) = \frac{p(1-p)e^{-u}}{\left(p + (1-p)e^{-u}\right)^2} \leq \frac{1}{4}.$$

By Taylor's theorem, $\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta)$ for some $\theta \in [0, u]$. Thus, we have:

$$\phi(u) = \phi(0) + u\phi'(0) + \frac{u^2}{2}\phi''(\theta) = \frac{u^2}{2}\phi''(\theta) \leq \frac{u^2}{8} = \frac{\lambda^2 (b-a)^2}{8}.$$

□

2.3.1 Understanding Hoeffding's Inequality

Hoeffding's inequality involves three interconnected terms: n, ε, and $\delta = 2e^{-2n\varepsilon^2}$, which is the bound on the probability that the event under $\mathbb{P}()$ holds (for the purpose of the discussion we consider two-sided Hoeffding's inequality for random variables bounded in $[0, 1]$). We can fix any two of the three terms n, ε, and δ and then the relation $\delta = e^{-2n\varepsilon^2}$ provides the value of the third. Thus, we have

$$\begin{aligned}
\delta &= 2e^{-2n\varepsilon^2}, \\
\varepsilon &= \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}, \\
n &= \frac{\ln\frac{2}{\delta}}{2\varepsilon^2}.
\end{aligned}$$

Overall, Hoeffding's inequality tells by how much the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$ can deviate from its expectation μ, but the interplay between the three parameters provides several ways of seeing and using Hoeffding's inequality. For example, if the number of samples n is fixed (we have made a fixed number of experiments and now analyze what we can get from them), there is an interplay between the precision ε and confidence δ. We can request higher precision ε, but then we have to compromise on the confidence δ that the desired bound $\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \leq \varepsilon$ holds. And the other way around: we can request higher confidence δ, but then we have to compromise on precision ε, i.e., we have to increase the allowed range $\pm\varepsilon$ around μ, where we expect to find the empirical average $\frac{1}{n}\sum_{i=1}^n X_i$.

As another example, we may have target precision ε and confidence δ and then the inequality provides us the number of experiments n that we have to perform in order to achieve the target.

It is often convenient to write the inequalities (2.4) and (2.5) with a fixed confidence in mind, thus we have

$$\begin{aligned}
\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) &\leq \delta, \\
\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^n X_i \geq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) &\leq \delta, \\
\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right| \geq \sqrt{\frac{\ln\frac{2}{\delta}}{2n}}\right) &\leq \delta.
\end{aligned}$$

(Put attention that the $\ln 2$ factor in the last inequality comes from the union bound over the first two inequalities: if we want to keep the same confidence we have to compromise on precision.)

In many situations we are interested in the complimentary events. Thus, for example, we have

$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \leq \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}\right) \geq 1 - \delta.$$

Careful reader may point out that the inequalities above should be strict ("$<$" and "$>$"). This is true, but if it holds for strict inequalities it also holds for non-strict inequalities ("$\leq$" and "$\geq$"). Since strict inequalities provide no practical advantage we will use the non-strict inequalities to avoid the headache of remembering which inequalities should be strict and which should not.

The last inequality essentially says that with probability at least $1 - \delta$ we have

$$\mu \leq \frac{1}{n}\sum_{i=1}^{n} X_i + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}$$

and this is how we will occasionally use it. Note that the random variable is $\frac{1}{n}\sum_{i=1}^{n} X_i$ and the right way of interpreting the above inequality is actually that with probability at least $1 - \delta$

$$\frac{1}{n}\sum_{i=1}^{n} X_i \geq \mu - \sqrt{\frac{\ln\frac{1}{\delta}}{2n}},$$

i.e., the probability is over $\frac{1}{n}\sum_{i=1}^{n} X_i$ and not over μ. However, many generalization bounds that we study in Chapter 3 are written in the first form in the literature and we follow the tradition.

2.4 Basics of Information Theory: Entropy, Relative Entropy, and the Method of Types

In this section we briefly introduce a number of basic concepts from information theory that are very useful for deriving concentration inequalities. Specifically, we introduce the notions of entropy and relative entropy (Cover and Thomas, 2006, Chapter 2) and some basic tools from the method of types (Cover and Thomas, 2006, Chapter 11). We start with some definitions.

Definition 2.7 (Entropy)**.** *Let $p(x)$ be a distribution of a discrete random variable X taking values in a finite set $\mathcal{X}$. We define the* entropy *of p as:*

$$\mathrm{H}(p) = -\sum_{x\in\mathcal{X}} p(x)\ln p(x).$$

We use the convention that $0\ln 0 = 0$ (which is justified by continuity of $z\ln z$, since $z\ln z \to 0$ as $z \to 0$).

We have special interest in Bernoulli random variables.

Definition 2.8 (Bernoulli random variable)**.** *X is a Bernoulli random variable with bias p if X accepts values in $\{0,1\}$ with $\mathbb{P}(X=0) = 1-p$ and $\mathbb{P}(X=1) = p$.*

Note that expectation of a Bernoulli random variable is equal to its bias:

$$\mathbb{E}[X] = 0 \times \mathbb{P}(X=0) + 1 \times \mathbb{P}(X=1) = \mathbb{P}(X=1) = p.$$

With a slight abuse of notation we specialize the definition of entropy to Bernoulli random variables.

Definition 2.9 (Binary entropy)**.** *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

$$\mathrm{H}(p) = -p\ln p - (1-p)\ln(1-p).$$

Note that when we talk about Bernoulli random variables p denotes the bias of the random variable and when we talk about more general random variables p denotes the complete distribution.

Entropy is one of the central quantities in information theory and it has numerous applications. We start by using binary entropy to bound binomial coefficients.

Lemma 2.10.

$$\frac{1}{n+1}e^{n\,\mathrm{H}\left(\frac{k}{n}\right)} \leq \binom{n}{k} \leq e^{n\,\mathrm{H}\left(\frac{k}{n}\right)}.$$

(Note that $\frac{k}{n} \in [0,1]$ and $\mathrm{H}\left(\frac{k}{n}\right)$ in the lemma is the binary entropy.)

Proof. By the binomial formula we know that for any $p \in [0,1]$:

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1. \tag{2.8}$$

We start with the upper bound. Take $p = \frac{k}{n}$. Since the sum is larger than any individual term, for the k-th term of the sum we get:

$$\begin{aligned}
1 &\geq \binom{n}{k} p^k \left(1-p\right)^{n-k} \\
&= \binom{n}{k} \left(\frac{k}{n}\right)^k \left(1-\frac{k}{n}\right)^{n-k} \\
&= \binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} \\
&= \binom{n}{k} e^{k \ln \frac{k}{n} + (n-k) \ln \frac{n-k}{n}} \\
&= \binom{n}{k} e^{n\left(\frac{k}{n} \ln \frac{k}{n} + \frac{n-k}{n} \ln \frac{n-k}{n}\right)} \\
&= \binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)}.
\end{aligned}$$

By changing sides of the inequality we obtain the upper bound.

For the lower bound it is possible to show that if we fix $p = \frac{k}{n}$ then $\binom{n}{k} p^k (1-p)^{n-k} \geq \binom{n}{i} p^i (1-p)^{n-i}$ for any $i \in \{0, \ldots, n\}$, see Cover and Thomas (2006, Example 11.1.3) for details. We also note that there are $n+1$ elements in the sum in equation (2.8). Again, take $p = \frac{k}{n}$, then

$$1 \leq (n+1) \max_i \binom{n}{i} \left(\frac{k}{n}\right)^i \left(\frac{n-k}{n}\right)^{n-i} = (n+1)\binom{n}{k} \left(\frac{k}{n}\right)^k \left(\frac{n-k}{n}\right)^{n-k} = (n+1)\binom{n}{k} e^{-n\,\mathrm{H}\left(\frac{k}{n}\right)},$$

where the last step follows the same steps as in the derivation of the upper bound. □

Lemma 2.10 shows that the number of configurations of chosing k out of n objects is directly related to the entropy of the imbalance $\frac{k}{n}$ between the number of objects that are selected (k) and the number of objects that are left out ($n-k$).

We now introduce one additional quantity, the *Kullback-Leibler (KL) divergence*, also known as *Kullback-Leibler distance* and as *relative entropy*.

Definition 2.11 (Relative entropy or Kullback-Leibler divergence)**.** *Let $p(x)$ and $q(x)$ be two probability distributions of a random variable X (or two probability density functions, if X is a continuous random variable), the* Kullback-Leibler divergence *or* relative entropy *is defined as:*

$$\mathrm{KL}(p\|q) = \mathbb{E}_p\left[\ln \frac{p(X)}{q(X)}\right] = \begin{cases} \sum_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)}, & \text{if } \mathcal{X} \text{ is discrete} \\ \int_{x\in\mathcal{X}} p(x) \ln \frac{p(x)}{q(x)} dx, & \text{if } \mathcal{X} \text{ is continuous} \end{cases}.$$

We use the convention that $0 \ln \frac{0}{0} = 0$ and $0 \ln \frac{0}{q} = 0$ and $p \ln \frac{p}{0} = \infty$.

We specialize the definition to Bernoulli distributions.

Definition 2.12 (Binary kl-divergence)**.** *Let p and q be biases of two Bernoulli random variables. The* binary kl divergence *is defined as:*

$$\mathrm{kl}(p\|q) = \mathrm{KL}([1-p, p]\|[1-q, q]) = p \ln \frac{p}{q} + (1-p) \ln \frac{1-p}{1-q}.$$

KL divergence is the central quantity in information theory. Although it is not a distance measure, because it does not satisfy the triangle inequality, it is the right way of measuring distances between probability distributions. This is illustrated by the following example.

Example 2.13. Let $X_1, \dots, X_n$ be an i.i.d. sample of n Bernoulli random variables with bias p and let $\frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias of the sample. (Note that $\frac{1}{n}\sum_{i=1}^n X_i \in \{0, \frac{1}{n}, \frac{2}{n}, \dots, \frac{n}{n}\}$.) Then by Lemma 2.10:

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) = \binom{n}{k} p^k (1-p)^{n-k} \leq e^{n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} = e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \tag{2.9}$$

and

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right) \geq \frac{1}{n+1} e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)}.$$

Thus, $\mathrm{kl}\left(\frac{k}{n}\|p\right)$ governs the probability of observing empirical bias $\frac{k}{n}$ when the true bias is p. It is easy to verify that $\mathrm{kl}(p\|p) = 0$ and it is also possible to show that $\mathrm{kl}(\hat{p}\|p)$ is convex in $\hat{p}$ and that $\mathrm{kl}(\hat{p}\|p) \geq 0$. Thus, the probability of empirical bias is maximized when it coincides with the true bias.

2.5 kl Inequality

Example 2.13 shows that kl can be used to bound the empirical bias when the true bias is known. But in machine learning we are usually interested in the inverse problem - how to infer the true bias p when the empirical bias $\hat{p}$ is known. Next we demonstrate that this is also possible and that it leads to an inequality, which in most cases is tighter than Hoeffding's inequality. We start with the following lemma.

Lemma 2.14. *Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias. Then*

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] \leq n+1.$$

Proof.

$$\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right] = \sum_{k=0}^n \mathbb{P}\left(\hat{p} = \frac{k}{n}\right) e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \leq \sum_{k=0}^n e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} e^{n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} = n+1,$$

where the inequality was derived in equation 2.9. □

We combine this lemma with Markov's inequality to obtain the following result.

Theorem 2.15 (kl inequality)**.** *Let $X_1, \dots, X_n$ be i.i.d. Bernoulli with bias p and let $\hat{p} = \frac{1}{n}\sum_{i=1}^n X_i$ be the empirical bias. Then*

$$\mathbb{P}(\mathrm{kl}(\hat{p}\|p) \geq \varepsilon) \leq (n+1)e^{-n\varepsilon}. \tag{2.10}$$

Proof. By Markov's inequality and Lemma 2.14:

$$\mathbb{P}(\mathrm{kl}(\hat{p}\|p) \geq \varepsilon) = \mathbb{P}\left(e^{n\,\mathrm{kl}(\hat{p}\|p)} \geq e^{n\varepsilon}\right) \leq \frac{\mathbb{E}\left[e^{n\,\mathrm{kl}(\hat{p}\|p)}\right]}{e^{n\varepsilon}} \leq \frac{n+1}{e^{n\varepsilon}}.$$

□

2.5.1 Relaxations of the kl-inequality: Pinsker's and refined Pinsker's inequalities

By denoting the right hand side of kl inequality (2.10) by δ, we obtain that with probability greater than $1-\delta$:

$$\mathrm{kl}(\hat{p}\|p) \leq \frac{\ln\frac{n+1}{\delta}}{n}. \tag{2.11}$$

This leads to an implicit bound on p, which is not very intuitive and not always convenient to work with. In order to understand the behavior of the kl inequality better we use a couple of its relaxations. The first relaxation is known as Pinsker's inequality, see Cover and Thomas (2006, Lemma 11.6.1).

Lemma 2.16 (Pinsker's inequality)**.**

$$\mathrm{KL}(p\|q) \geq \frac{1}{2}\|p-q\|_1^2,$$

where $\|p-q\|_1 = \sum_{x\in\mathcal{X}} |p(x)-q(x)|$ *is the* L_1*-norm.*

Corollary 2.17 (Pinsker's inequality for the binary kl divergence)**.**

$$\mathrm{kl}(p\|q) \geq \frac{1}{2}\left(|p-q| + |(1-p)-(1-q)|\right)^2 = 2(p-q)^2. \tag{2.12}$$

By applying Corollary 2.17 to inequality (2.11) we obtain that with probability greater than $1-\delta$

$$|p-\hat{p}| \leq \sqrt{\frac{\mathrm{kl}(\hat{p}\|p)}{2}} \leq \sqrt{\frac{\ln\frac{n+1}{\delta}}{2n}}.$$

Recall that Hoeffding's inequality assures that with probability greater than $1-\delta$

$$p \leq \hat{p} + \sqrt{\frac{\ln\frac{1}{\delta}}{2n}}.$$

Thus, in the worst case the kl inequality is only weaker by the $\ln(n+1)$ factor and in fact the $\ln(n+1)$ factor can be reduced by a more careful analysis, see Maurer (2004), Langford (2005). Next we show that the kl inequality can actually be significantly tighter than Hoeffding's inequality. For this we use refined Pinsker's inequality, see Marton (1996, 1997), Samson (2000), Boucheron et al. (2013, Lemma 8.4).

Lemma 2.18 (Refined Pinsker's inequality)**.**

$$\mathrm{kl}(p\|q) \geq \frac{(p-q)^2}{2\max\{p,q\}} + \frac{(p-q)^2}{2\max\{(1-p),(1-q)\}}.$$

Corollary 2.19 (Refined Pinsker's inequality)**.** *If* $q > p$ *then*

$$\mathrm{kl}(p\|q) \geq \frac{(p-q)^2}{2q}.$$

Corollary 2.20 (Refined Pinsker's inequality)**.** *If* $\mathrm{kl}(p\|q) \leq \varepsilon$ *then*

$$q \leq p + \sqrt{2p\varepsilon} + 2\varepsilon.$$

By applying Corollary 2.20 to inequality (2.11) we obtain that with probability greater than $1-\delta$

$$p \leq \hat{p} + \sqrt{\frac{2\hat{p}\ln\frac{n+1}{\delta}}{n}} + \frac{2\ln\frac{n+1}{\delta}}{n}.$$

Note that when $\hat{p}$ is close to zero, the latter inequality is much tighter than Hoeffding's inequality. Finally, we note that although there is no analytic inversion of $\mathrm{kl}(\hat{p}\|p)$ it is possible to invert it numerically to obtain even tighter bounds than the relaxations above. Additionally, the bound in Theorem 2.15 can be improved slightly, see Maurer (2004), Langford (2005).

2.6 Sampling Without Replacement

Let $X_1,\ldots,X_n$ be a sequence of random variables *sampled without replacement* from a finite set of values $\mathcal{X} = \{x_1,\ldots,x_N\}$ of size N. The random variables $X_1,\ldots,X_n$ are *dependent*. For example, if $\mathcal{X} = \{-1,+1\}$ and we sample two values then $X_1 = -X_2$. Since $X_1,\ldots,X_n$ are dependent, the concentration results from previous sections do not apply directly. However, the following result by Hoeffding (1963, Theorem 4), which we cite without a proof, allows to extend results for sampling with replacement to sampling without replacement.

Lemma 2.21. *Let $X_1, \ldots, X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of N real values. Let $Y_1, \ldots, Y_n$ denote a random sample with replacement from $\mathcal{X}$. Then for any continuous and convex function $f : \mathbb{R} \to \mathbb{R}$*

$$\mathbb{E}\left[f\left(\sum_{i=1}^{n} X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^{n} Y_i\right)\right].$$

In particular, the lemma can be used to prove Hoeffding's inequality for sampling without replacement.

Theorem 2.22 (Hoeffding's inequality for sampling without replacement)**.** *Let $X_1, \ldots, X_n$ denote a random sample without replacement from a finite set $\mathcal{X} = \{x_1, \ldots, x_N\}$ of N values, where each element x_i is in the $[0, 1]$ interval. Let $\mu = \frac{1}{N}\sum_{i=1}^{N} x_i$ be the average of the values in $\mathcal{X}$. Then for all $\varepsilon > 0$*

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu \geq \varepsilon\right) \leq e^{-2n\varepsilon^2},$$
$$\mathbb{P}\left(\mu - \frac{1}{n}\sum_{i=1}^{n} X_i \geq \varepsilon\right) \leq e^{-2n\varepsilon^2}.$$

The proof is a minor adaptation of the proof of Hoeffding's inequality for sampling with replacement using Lemma 2.21 and is left as an exercise. (Note that it requires a small modification inside the proof, because Lemma 2.21 cannot be applied directly to the statement of Hoeffding's inequality.)

While formal proof requires a bit of work, intuitively the result is quite expected. Imagine the process of sampling without replacement. If the average of points sampled so far starts deviating from the mean of the values in $\mathcal{X}$, the average of points that are left in $\mathcal{X}$ deviates in the opposite direction and "applies extra force" to new samples to bring the average back to μ. In the limit when $n = N$ we are guaranteed to have the average of X_i-s being equal to μ.

Chapter 3

Generalization Bounds for Classification

One of the most central questions in machine learning is: "How much can we trust the predictions of a learning algorithm?". A way of answering this question is by providing generalization bounds on the expected performance of the algorithm on new data points. In this chapter we derive a number of generalization bounds for supervised classification.

3.1 Overview: Learning by Selection

The classical process of learning can be seen as a selection process (see Figure 3.1):

1. We start with a hypothesis set $\mathcal{H}$, which is a set of plausible prediction rules (for example, linear separators).
2. We observe a sample S sampled i.i.d. according to a fixed, but unknown distribution $p(X, Y)$.
3. Based on the empirical performances $\hat{L}(h, S)$ of the hypotheses in $\mathcal{H}$, we *select* a prediction rule $\hat{h}_S^*$, which we consider to be the "best" in $\mathcal{H}$ in some sense. Typically, $\hat{h}_S^*$ is either the *empirical risk minimizer* (ERM), $\hat{h}_S^* = \arg\min_h \hat{L}(h, S)$, or a regularized empirical risk minimizer.
4. $\hat{h}_S^*$ is then applied to predict labels for new samples X.

In this chapter we are concerned with the question of what can be said about the expected loss $L(\hat{h}_S^*)$, which is the error we are expected to make on new samples. More precisely, we provide tools for bounding the probability that $\hat{L}(\hat{h}_S^*, S)$ is significantly smaller than $L(\hat{h}_S^*)$. Recall that $\hat{L}(\hat{h}_S^*, S)$ is observed and $L(\hat{h}_S^*)$ is unobserved. Having small $\hat{L}(\hat{h}_S^*, S)$ and large $L(\hat{h}_S^*)$ is undesired, because it means that based on $\hat{L}(\hat{h}_S^*, S)$ we believe that $\hat{h}_S^*$ performs well, but in reality it does not.

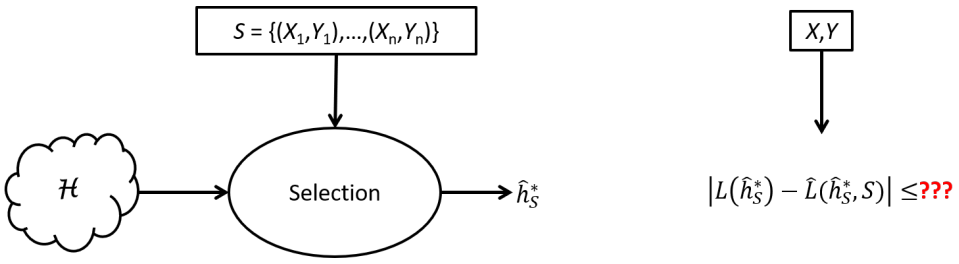


Figure 3.1: Learning by Selection.

Assumtions ***There are two key assumptions we make throughout the chapter:***

1. ***The samples in S are i.i.d..***
2. ***The new samples (X, Y) come from the same distribution as the samples in S.***

These are the assumptions behind concentration of measure inequalities developed in Chapter 2 and it is important to remember that if they are not satisfied the results derived in this chapter are not valid.

In a sense, it is intuitive why we have to make these assumptions. For example, if we train a language model using data from The Wall Street Journal and then apply it to Twitter the change in prediction accuracy can be very dramatic. Even though both are written in English and comprehensible by humans, the language used by professional journalists writing for The Wall Street Journal is very different from the language used in the short tweets.

The two assumptions are behind most supervised learning algorithms that you can meet in practice and, therefore, it is important to keep them in mind. In Chapter 5 we discuss how to departure from them, but for now we stick with them.

Given the assumptions above, for any fixed prediction rule that is independent of S, the empirical loss is an unbiased estimate of the true loss, $\mathbb{E}\left[\hat{L}(h, S)\right] = L(h)$. An intuitive way to see it is that under the assumptions that the samples in S are i.i.d. and coming from the same distribution as new samples (X, Y), from the perspective of h the new samples (X, Y) are in no way different from the samples in S: any new sample (X, Y) could have happened to be in S instead of some other sample (X_i, Y_i) (they are "exchangeable"). Formally,

$$\begin{aligned}
\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\hat{L}(h, S)\right] &= \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(h(X_i), Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\ell(h(X_i), Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(X_i,Y_i)}\left[\ell(h(X_i), Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}L(h) \\
&= L(h).
\end{aligned}$$

However, when we make the selection of $\hat{h}^*_S$ based on S the "exchangeability" argument no longer applies and $\mathbb{E}\left[\hat{L}(\hat{h}^*_S, S)\right] \neq \mathbb{E}\left[L(\hat{h}^*_S)\right]$ (note that $\hat{h}^*_S$ is a random variable depending on S and we take expectation with respect to this randomness). This is because $\hat{h}^*_S$ is tailored to S (for example, it minimizes $\hat{L}(h, S)$) and from the perspective of selection process the samples in S are not exchangeable with new samples (X, Y). If we exchange the samples we may end up with a different $\hat{h}^*_S$. In the extreme case when the hypothesis space $\mathcal{H}$ is so rich that it can fit any possible labeling of the data (for example, the hypothesis space corresponding to 1-nearest-neighbor prediction rule) we may end up in a situation, where $\hat{L}(\hat{h}^*_S, S)$ is always zero, but $\mathbb{E}\left[L(\hat{h}^*_S)\right] \geq \frac{1}{4}$, as in the following informal example.

Informal Lower Bound Imagine that we want to learn a classifier that predicts whether a student's birthday is on an even or odd day based on student's id. Assume that the total number of students is $2n$, that the hypothesis class $\mathcal{H}$ includes all possible mappings from student id to even/odd, so that $|\mathcal{H}| = 2^{2n}$, and that we observe a sample of n uniformly sampled students (potentially with repetitions). Since all possible mappings are within $\mathcal{H}$, we have $\hat{h}^*_S \in \mathcal{H}$ for which $\hat{L}(\hat{h}^*_S, S) = 0$. However, $\hat{h}^*_S$ is guaranteed to make zero error only on the samples that were observed, which constitute at most half of the total number of students. For the remaining students $\hat{h}^*_S$ can, at the best, make a random guess which will succeed with probability $\frac{1}{2}$. Therefore, the expected loss of $\hat{h}^*_S$ is $L(\hat{h}^*_S) \geq \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$, where the first term is an upper bound on the probability of observing an already seen student times the expected error $\hat{h}^*_S$ makes in this case and the second term is a lower bound on the probability of

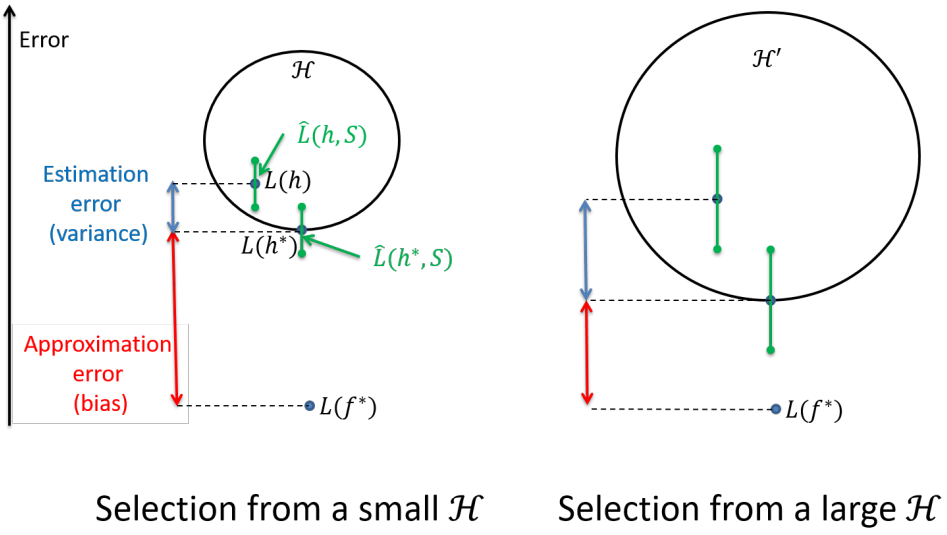


Figure 3.2: Learning by Selection.

observing a new student times the expected error $\hat{h}_S^*$ makes in this case. For a more formal treatment see the lower bounds in Chapter 3.7.

Considering it from the perspective of expectations, we have:

$$\begin{aligned}
\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\hat{L}(\hat{h}_S^*,S)\right] &= \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\hat{h}_S^*(X_i),Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\ell(\hat{h}_S^*(X_i),Y_i)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\ell(\hat{h}_S^*(X_1),Y_1)\right] \\
&= \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\ell(\hat{h}_S^*(X_1),Y_1)\right] \\
&\neq \mathbb{E}_{(X,Y)}\left[\mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\ell(\hat{h}_S^*(X),Y)\right]\right] \\
&= \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[\mathbb{E}_{(X,Y)}\left[\ell(\hat{h}_S^*(X),Y)\right]\right] \\
&= \mathbb{E}_{(X_1,Y_1),\ldots,(X_n,Y_n)}\left[L(\hat{h}_S^*)\right].
\end{aligned}$$

The selection leads to the approximation-estimation trade-off (a.k.a. bias-variance trade-off), see Figure 3.2. If the hypothesis class $\mathcal{H}$ is small it is easy to identify a good hypothesis h in $\mathcal{H}$, but since $\mathcal{H}$ is small it is likely that all the hypotheses in $\mathcal{H}$ are weak. On the other hand, if $\mathcal{H}$ is large it is more likely to contain stronger hypotheses, but at the same time the probability of confusion with a poor hypothesis grows. This is because there is always a small chance that the empirical loss $\hat{L}(h,S)$ does not represent the true loss $L(h)$ faithfully. The more hypotheses we take, the higher is the chance that $\hat{L}(h,S)$ is misleading for some of them, which increases the chance of confusion.

Finding a good balance between approximation and estimation errors is one of the central questions in machine learning. The main tool for analyzing the trade-off from the theoretical perspective are concentration of measure inequalities. Since concentration of measure inequalities do not apply when the prediction rule $\hat{h}_S^*$ depends on S, the main approach to analyzing the prediction power of $\hat{h}_S^*$ is to consider cases with no dependency and then take a union bound over selection from these cases. In this chapter we study three different ways of implementing this idea, see Figure 3.3 for an overview. We distinguish between *hard selection*, where the learning procedure returns a single hypothesis h and *soft selection*, where the learning procedure returns a distribution over $\mathcal{H}$.

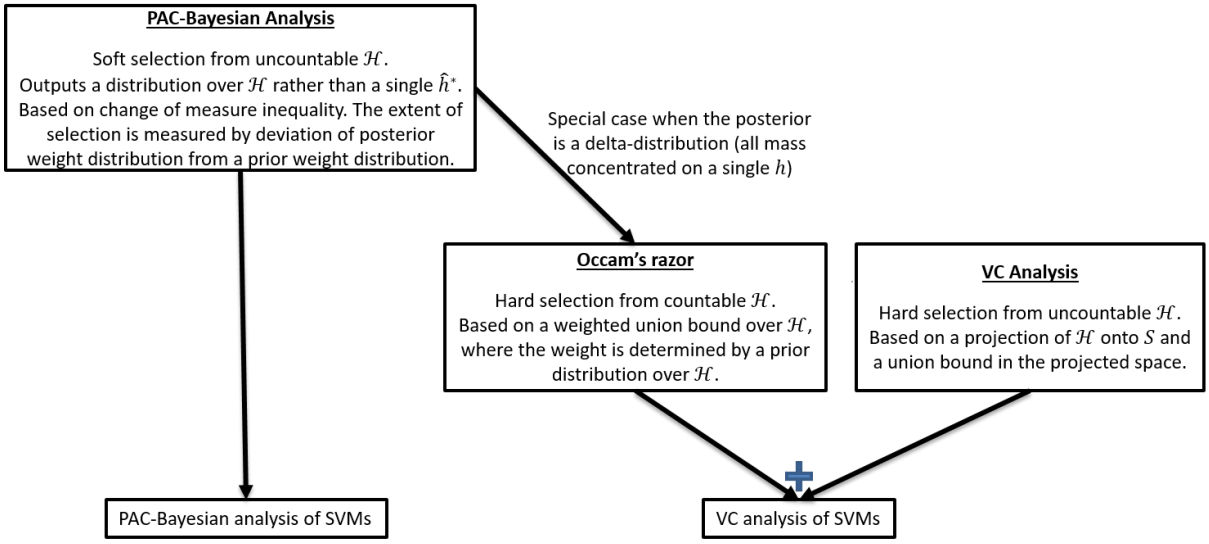


Figure 3.3: Overview of the major approaches to derivation of generalization bounds considered in this chapter.

1. *Occam's razor* applies to *hard selection* from a *countable* hypothesis space $\mathcal{H}$ and it is based on a weighted union bound over $\mathcal{H}$. We know that for every fixed h the expected loss is close to the empirical loss, meaning that $\left|L(h) - \hat{L}(h, S)\right|$ is small. When $\mathcal{H}$ is countable we can take a weighted union bound and obtain that $\left|L(h) - \hat{L}(h, S)\right|$ is "small" for all $h \in \mathcal{H}$ (where the magnitude of "small" is inversely proportional to the weight of h in the union bound) and thus it is "small" for $\hat{h}^*_S$.

2. *Vapnik-Chervonenkis (VC) analysis* applies to *hard selection* from an *uncountable* hypothesis space $\mathcal{H}$ and it is based on projection of $\mathcal{H}$ onto S and a union bound over what we obtain after the projection. The idea is that even when $\mathcal{H}$ is uncountably infinite, there is only a finite number of "behaviors" (ways to label S) we can observe on a finite sample S. In other words, when we look at $\mathcal{H}$ through the prism of S we can only distinguish between a finite number of subsets of $\mathcal{H}$ and everything that falls within the subsets is equivalent in terms of $\hat{L}(h, S)$. Therefore, S only serves for a (finite) selection of a subset of $\mathcal{H}$ out of a finite number of subsets, whereas the (infinite) selection from within the subset is independent of S. Selection that is independent of S introduces no bias. As before, the VC analysis exploits the fact that for any fixed h the distance $\left|L(h) - \hat{L}(h, S)\right|$ is small and then takes a union bound over the potential dependencies, which are the dependencies between the subsets (the projections) and S.

3. *PAC-Bayesian analysis* applies to *soft selection* from an *uncountable* hypothesis space $\mathcal{H}$ and it is based on *change of measure inequality*, which can be seen as a refinement of the union bound. Unlike the preceding two approaches, which return a single classifier $\hat{h}^*_S$, PAC-Bayesian analysis returns a *randomized classifier* defined by a distribution ρ over $\mathcal{H}$. The actual classification then happens by drawing a new classifier h from $\mathcal{H}$ according to ρ at each prediction round and applying it to make a prediction. When $\mathcal{H}$ is countable, ρ can (but does not have to) be a delta-distribution putting all the mass on a single hypothesis $\hat{h}^*_S$ and in this case the generalization guarantees are identical to those in Occam's razor approach. The amount of selection is measured by deviation of ρ from a prior distribution π, where π is selected independently of S. It is natural to put more of ρ-mass on hypotheses that perform well on S, but the more we skew ρ toward well-performing hypotheses the more it deviates from π. This provides a more refined way of measuring the amount of selection compared to the other two approaches. Furthermore, randomization allows to avoid selection when it is not necessary. The avoidance of selection reduces the variance without impairing the bias. For example, when two hypotheses have similar empirical performance we do not have

to commit to one of them, but can instead distribute ρ equally among them. The analysis then provides a certain "bonus" for avoiding commitment.

3.2 Generalization Bound for a Single Hypothesis

We start with the simplest case, where $\mathcal{H}$ consists of a single prediction rule h. We are interested in the quality of h, measured by $L(h)$, but all we can measure is $\hat{L}(h, S)$. What can we say about $L(h)$ based on $\hat{L}(h, S)$? Note that the samples $(X_i, Y_i) \in S$ come from the same distribution as any future samples (X, Y) we will observe. Therefore, $\ell(h(X_i), Y_i)$ has the same distribution as $\ell(h(X), Y)$ for any future sample (X, Y). Let $Z_i = \ell(h(X_i), Y_i)$ be the loss of h on (X_i, Y_i). Then $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} Z_i$ is an average of n i.i.d. random variables with $\mathbb{E}[Z_i] = \mathbb{E}[\ell(h(X), Y)] = L(h)$. The distance between $\hat{L}(h, S)$ and $L(h)$ can thus be bounded by application of Hoeffding's inequality.

Theorem 3.1. *Assume that ℓ is bounded in the $[0, 1]$ interval (i.e., $\ell(Y', Y) \in [0, 1]$ for all Y', Y), then for a single h and any $\delta \in (0, 1)$ we have:*

$$\mathbb{P}\left(L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}\right) \leq \delta \tag{3.1}$$

and

$$\mathbb{P}\left(\left|L(h) - \hat{L}(h, S)\right| \geq \sqrt{\frac{\ln \frac{2}{\delta}}{2n}}\right) \leq \delta. \tag{3.2}$$

Proof. For (3.1) take $\varepsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$ in (2.5) and rearrange the terms. Equation (3.2) follows in a similar way from the two-sided Hoeffding's inequality. Note that in (3.1) we have $\frac{1}{\delta}$ and in (3.2) we have $\frac{2}{\delta}$. □

There is an alternative way to read equation (3.1): with probability at least $1 - \delta$ we have

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}.$$

We remind the reader that the above inequality should actually be interpreted as

$$\hat{L}(h, S) \geq L(h) - \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$$

and it means that with probability at least $1 - \delta$ the empirical loss $\hat{L}(h, S)$ does not underestimate the expected loss $L(h)$ by more than $\sqrt{\ln(1/\delta)/2n}$. However, it is customary to write the inequality in the first form (as an upper bound on $L(h)$ and we follow the tradition (see the discussion at the end of Section 2.3.1).

Theorem 3.1 is analogous to the problem of estimating a bias of a coin based on coin flip outcomes. There is always a small probability that the flip outcomes will not be representative of the coin bias. For example, it may happen that we flip a fair coin 1000 times (without knowing that it is a fair coin!) and observe "all heads" or some other misleading outcome. And if this happens we are doomed - there is nothing we can do when the sample does not represent the reality faithfully. Fortunately for us, this happens with a small probability that decreases exponentially with the sample size n.

Whether we use the one-sided bound (3.1) or the two-sided bound (3.2) depends on the situation. In most cases we are interested in the upper bound on the expected performance of the prediction rule given by (3.1).

3.3 Generalization Bound for Finite Hypothesis Classes

A hypothesis set $\mathcal{H}$ containing a single hypothesis is a very boring set. In fact, we cannot learn in this case, because we end up with the same single hypothesis no matter what the sample S is. Learning

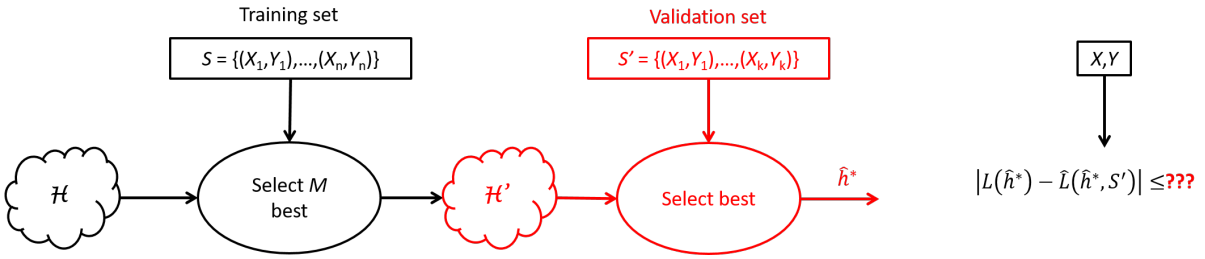


Figure 3.4: Validation (the red part in the figure) is identical to learning with a reduced hypothesis set $\mathcal{H}'$ (most often $\mathcal{H}'$ is finite).

becomes interesting when training sample S helps to improve future predictions or, equivalently, decrease the expected loss $L(h)$. In this section we consider the simplest non-trivial case, where $\mathcal{H}$ consists of a finite number of hypotheses M. There are at least two cases, where we meet a finite $\mathcal{H}$ in real life. The first is when the input space $\mathcal{X}$ is finite. This case is relatively rare. The second and much more frequent case is when $\mathcal{H}$ itself is an outcome of a learning process. For example, this is what happens in a validation procedure, see Figure 3.4. In validation we are using a validation set in order to select the best hypothesis out of a finite number of candidates corresponding to different parameter values and/or different algorithms.

And now comes the delicate point. Let $\hat{h}_S^*$ be a hypothesis with minimal empirical risk, $\hat{h}_S^* = \arg\min_h \hat{L}(h, S)$ (it is natural to pick the empirical risk minimizer $\hat{h}_S^*$ to make predictions on new samples, but the following discussion equally applies to any other selection rule that takes sample S into account; note that there may be multiple hypotheses that achieve the minimal empirical error and in this case we can pick one arbitrarily). While for each h individually $\mathbb{E}\left[\hat{L}(h, S)\right] = L(h)$, this is not true for $\mathbb{E}\left[\hat{L}(\hat{h}_S^*, S)\right]$. In other words, $\mathbb{E}\left[\hat{L}(\hat{h}_S^*, S)\right] \neq \mathbb{E}\left[L(\hat{h}_S^*)\right]$ (we have to put expectation on the right hand side, because $\hat{h}_S^*$ depends on the sample). The reason is that when we pick $\hat{h}_S^*$ that minimizes the empirical error on S, from the perspective of $\hat{h}_S^*$ the samples in S no longer look identical to future samples (X, Y). This is because $\hat{h}_S^*$ is selected in a very special way - it is selected to minimize the empirical error on S and, thus, it is tailored to S and most likely does better on S than on new random samples (X, Y). One way to handle this issue is to apply a union bound.

Theorem 3.2. *Assume that ℓ is bounded in the $[0, 1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0, 1)$ we have:*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right) \leq \delta. \tag{3.3}$$

Proof.

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right) \leq \sum_{h \in \mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}\right) \leq \sum_{h \in \mathcal{H}} \frac{\delta}{M} = \delta,$$

where the first inequality is by the union bound and the second is by Hoeffding's inequality. □

Another way of reading Theorem 3.2 is: with probability at least $1 - \delta$ for all $h \in \mathcal{H}$

$$L(h) \leq \hat{L}(h, S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}. \tag{3.4}$$

It means that no matter which h from $\mathcal{H}$ is returned by the algorithm, with high probability we have the guarantee (3.4). In particular, it holds for $\hat{h}_S^*$. Again, remember that the random quantity is actually $\hat{L}(h, S)$ and the right way to read the bound is $\hat{L}(h, S) \geq L(h) - \sqrt{\ln(M/\delta)/2n}$, see the discussion in the previous section.

The price for considering M hypotheses instead of a single one is $\ln M$. Note that it grows only logarithmically with M! Also note that there is no contradiction between the upper bound and the lower bound we have discussed in Section 3.1. In the construction of the lower bound we took $M = |\mathcal{H}| = 2^{2n}$. If we substitute this value of M into (3.4) we obtain $\sqrt{\ln(M/\delta)/2n} \geq \sqrt{\ln(2)} \geq 0.8$, which has no contradiction with $L(h) \geq 0.25$.

Similar to theorem 3.1 it is possible to derive a two-sided bound on the error. It is also possible to derive a lower bound by using the other side of Hoeffding's inequality (2.4): with probability at least $1-\delta$, for all $h \in \mathcal{H}$ we have $L(h) \geq \hat{L}(h,S) - \sqrt{\ln(M/\delta)/2n}$. Typically we want the upper bound on $L(h)$, but if we want to compare two prediction rules, h and h', we need an upper bound for one and a lower bound for the other. The "lazy" approach is to take the two-sided bound for everything, but sometimes it is possible to save the factor of $\ln(2)$ by carefully considering which hypotheses require the lower bound and which require the upper bound and applying the union bound correspondingly (we are not getting into the details).

3.4 Occam's Razor Bound

Now we take a closer look at Hoeffding's inequality. It says that

$$\mathbb{P}\left(L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\delta}\right)}{2n}} \right) \leq \delta,$$

where δ is the probability that things go wrong and $\hat{L}(h,S)$ happens to be far away from $L(h)$ because S is not representative for the performance of h. There is a dependence between the probability that things go wrong and the requirement on the closeness between $L(h)$ and $\hat{L}(h,S)$. If we want them to be very close (meaning that $\ln\left(\frac{1}{\delta}\right)$ is small) then δ has to be large, but if we can allow larger distance then δ can be smaller.

So, δ can be seen as our "confidence budget" (or, more precisely, "uncertainty budget") - the probability that we allow things to go wrong. The idea behind Occam's Razor bound is to distribute this budget unevenly among the hypotheses in $\mathcal{H}$. We use $\pi(h) \geq 0$, such that $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$ as our distribution of the confidence budget δ, where each hypothesis h is assigned $\pi(h)$ fraction of the budget. This means that for every hypothesis $h \in \mathcal{H}$ the sample S is allowed to be "non representative" with probability at most $\pi(h)\delta$, so that the probability that there exists any $h \in \mathcal{H}$ for which S is not representative is at most δ (by the union bound). The price that we pay is that the precision (the closeness of $\hat{L}(h,S)$ to $L(h)$) now differs from one hypothesis to another and depends on the confidence budget $\pi(h)\delta$ that was assigned to it. More precisely, $\hat{L}(h,S)$ is allowed to underestimate $L(h)$ by up to $\sqrt{\ln(1/(\pi(h)\delta))/2n}$. The precision increases when $\pi(h)$ increases, but since $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$ we cannot afford high precision for every h and have to compromise. More on this in the next theorem and its applications that follow.

Theorem 3.3. *Let ℓ be bounded in $[0,1]$, let $\mathcal{H}$ be a countable hypothesis set and let $\pi(h)$ be independent of the sample and satisfying $\pi(h) \geq 0$ for all h and $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$. Then:*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}} \right) \leq \delta.$$

Proof.

$$\begin{aligned}
\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}} \right) &\leq \sum_{h\in\mathcal{H}} \mathbb{P}\left(L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}} \right) \\
&\leq \sum_{h\in\mathcal{H}} \pi(h)\delta \\
&\leq \delta,
\end{aligned}$$

where the first inequality is by the union bound, the second inequality is by Hoeffding's inequality, and the last inequality is by the assumption on $\pi(h)$. Note that $\pi(h)$ has to be selected before we observe the sample (or, in other words, independently of the sample), otherwise the second inequality does not hold. More explicitly, in Hoeffding's inequality $\mathbb{P}\Big(\mathbb{E}\left[Z_1\right] - \frac{1}{n}\sum_{i=1}^{n} Z_i \geq \sqrt{\ln(1/\delta')/2n}\Big) \leq \delta'$ the parameter δ' has to be independent of $Z_1, \ldots, Z_n$. For $\pi(h)$ independent of S we take $\delta' = \pi(h)\delta$ and apply the inequality. But if $\pi(h)$ would be dependent on S we would not be able to apply it. □

Another way of reading Theorem 3.3 is that with probability at least $1-\delta$, for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln\left(\frac{1}{\pi(h)\delta}\right)}{2n}}.$$

Again, refer back to the discussion in Section 3.2 regarding the correct interpretation of the inequality. Note that the bound on $L(h)$ depends both on $\hat{L}(h,S)$ and on $\pi(h)$. Therefore, according to the bound, the best generalization is achieved by h that optimizes the trade-off between empirical performance $\hat{L}(h,S)$ and $\pi(h)$, where $\pi(h)$ can be interpreted as a complexity measure or a prior belief. Also, note that $\pi(h)$ can be designed arbitrarily, but it should be independent of the sample S. If $\pi(h)$ happens to put more mass on h-s with low $\hat{L}(h,S)$ the bound will be tighter, otherwise the bound will be loser, but it will still be a valid bound. But we cannot readjust $\pi(h)$ after observing S! Some considerations behind the choice of $\pi(h)$ are provided in Section 3.4.1.

Also note that while we can select $\pi(h)$ such that $\sum_{h\in\mathcal{H}} \pi(h) = 1$ and interpret π as a probability distribution over $\mathcal{H}$, it is not a requirement (we may have $\sum_{h\in\mathcal{H}} \pi(h) < 1$) and π is used as an auxiliary construction for derivation of the bound rather than the prior distribution in the Bayesian sense (for readers who are familiar with Bayesian learning). However, we can use π to incorporate prior knowledge into the learning procedure.

3.4.1 Applications of Occam's Razor bound

We consider two applications of Occam's Razor bound.

Generalization bound for finite hypotheses spaces

An immediate corollary of Occam's razor bound is the generalization bound for finite hypotheses classes that we have already seen in Section 3.3.

Corollary 3.4. *Let $\mathcal{H}$ be a finite hypotheses class of size M, then*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(M/\delta\right)}{2n}}\right) \leq \delta.$$

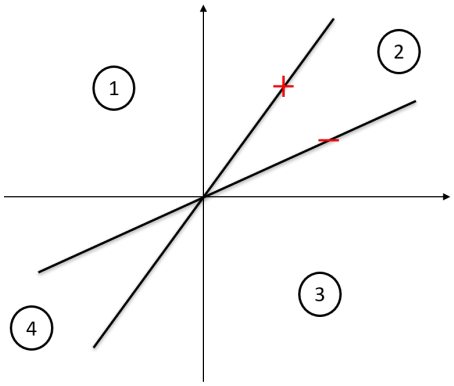
Proof. We set $\pi(h) = \frac{1}{M}$ (which means that we distribute the confidence budget δ uniformly among the hypotheses in $\mathcal{H}$) and apply Theorem 3.3. □

Generalization bound for binary decision trees

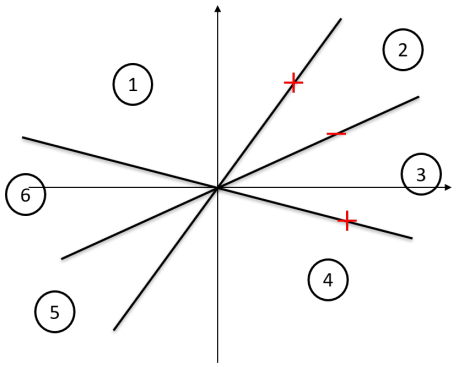
Theorem 3.5. *Let $\mathcal{H}_d$ be the set of binary decision trees of depth d and let $\mathcal{H} = \bigcup_{d=0}^{\infty} \mathcal{H}_d$ be the set of binary decision trees of unlimited depth. Let $d(h)$ be the depth of tree (hypothesis) h. Then*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h,S) + \sqrt{\frac{\ln\left(2^{2^{d(h)}} 2^{d(h)+1}/\delta\right)}{2n}}\right) \leq \delta.$$

Proof. We first note that $|\mathcal{H}_d| = 2^{2^d}$. We define $\pi(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{2^{d(h)}}}$. The first part of $\pi(h)$ distributes the confidence budget δ among $\mathcal{H}_d$-s (we can see it as $p(\mathcal{H}_d) = \frac{1}{2^{d(h)+1}}$, the share of confidence budget that goes to $\mathcal{H}_d$) and the second part of $\pi(h)$ distributes the confidence budget uniformly within $\mathcal{H}_d$. Since $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$, the assumption $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$ is satisfied. The result follows by application of Theorem 3.3. □



(a) Subsets of linear homogeneous separators defined by two sample points.



(b) Subsets of linear homogeneous separators defined by three sample points.

Figure 3.5: **Subsets of homogeneous linear separators in $\mathbb{R}^2$ formed by 3.5a two and 3.5b three sample points.** A homogeneous linear separator in $\mathbb{R}^2$ is defined by a vector $w \in \mathbb{R}^2$. The sample points define a number of regions in $\mathbb{R}^2$ that are shown by the numbers in circles. We say that a linear separator falls within a certain region when the vector w defining it falls within that region. All homogeneous linear separators falling within the same region have the same empirical loss $\hat{L}(h, S)$ and, therefore, any selection among them is not based on the sample S and introduces no bias. The sample only discriminates between the subsets.

Note that the bound depends on $\ln\left(\frac{1}{\pi(h)\delta}\right)$ and the dominating term in $\frac{1}{\pi(h)}$ is $2^{2^{d(h)}}$. We could have selected a different distribution of confidence over $\mathcal{H}_d$-s, for example, $p(\mathcal{H}_d) = \frac{1}{(d+1)(d+2)}$ (for which we also have $\sum_{d=0}^{\infty} \frac{1}{(d+1)(d+2)} = 1$), which is perfectly fine, but makes no significant difference for the bound. The dominating complexity term $\ln\left(2^{2^{d(h)}}\right)$ comes from uniform distribution of confidence within $\mathcal{H}_d$, which makes sense unless we have some prior information about the problem. In absence of such information there is no reason to give preference to any of the trees within $\mathcal{H}_d$, because $\mathcal{H}_d$ is symmetric.

The prior selected in the proof of Theorem 3.5 exploits structural symmetries within the hypothesis class $\mathcal{H}$ and assigns equal weight to hypotheses that are symmetric under permutation of names of the input variables. While we want $\pi(h)$ to be as large as possible for every h, the number of such permutation symmetric hypotheses is the major barrier dictating how large $\pi(h)$ can be (because π has to satisfy $\sum_{h\in\mathcal{H}} \pi(h) \leq 1$). Deeper trees have more symmetric permutations and, therefore, get smaller $\pi(h)$ compared to shallow trees. If there is prior information that breaks the permutation symmetry it can be used to assign higher prior to the corresponding trees and if it correctly reflects the true data distribution it will also lead to tighter bounds. If the prior information does not match the true data distribution such adjustments may have the opposite effect.

3.5 Vapnik-Chervonenkis (VC) Analysis

Now we present Vapnik-Chervonenkis (VC) analysis of generalization when a hypothesis is selected from an uncountably infinite hypothesis class $\mathcal{H}$. The reason that we are able to derive a generalization bound even though we are selecting from an uncountably large set is that only a finite part of this selection is based on the sample S, whereas the remaining uncountable selection is not based on the sample and, therefore, introduces no bias. Since the sample is finite, the number of distinct labeling patterns, also called dichotomies, $(h(X_1), \ldots, h(X_n))$ is also finite. When two hypotheses, h and h', produce the same labeling pattern, $(h(X_1), \ldots, h(X_n)) = (h'(X_1), \ldots, h'(X_n))$, the sample does not discriminate between them and the selection between h and h' is based on some other considerations rather than the sample. Therefore, the sample defines a finite number of (typically uncountably infinite) subsets of the hypothesis space $\mathcal{H}$, where hypotheses within the same subset produce the same labeling pattern $(h(X_1), \ldots, h(X_n))$. The sample then allows selection of the "best" subset, for example, the subset that minimizes the empirical error. All prediction rules within the same subset have the same empirical error

$\hat{L}(h, S)$ and selection among them is independent of S. See Figure 3.5 for an illustration.

The *effective selection* based on the sample S depends on the number of subsets of $\mathcal{H}$ with distinct labeling patterns on S. When the number of such subsets is exponential in the size of the sample n, the selection is too large and leads to overfitting, as we have already seen for selection from large finite hypothesis spaces in the earlier sections. I.e., we cannot guarantee closeness of $\hat{L}(\hat{h}^*_S, S)$ to $L(\hat{h}^*_S)$. However, if the number of subsets is subexponential in n, we can provide generalization guarantees for $L(\hat{h}^*_S)$. In Figure 3.5 we illustrate (informally) that at a certain point the number of subsets of the class of homogeneous linear separators in $\mathbb{R}^2$ stops growing exponentially with n.[1] For $n = 2$ the sample defines $4 = 2^n$ subsets, but for $n = 3$ the sample defines $6 < 2^n$ subsets. It can be formally shown that no 3 sample points can define more than 6 subsets of the space of homogeneous linear separators in $\mathbb{R}^2$ (some may define less, but that is even better for us) and that for $n > 2$ the number of subsets grows polynomially rather than exponentially with n.

In what follows we first bound the distance between $\hat{L}(h, S)$ and $L(h)$ for all $h \in \mathcal{H}$ in terms of the number of subsets using symmetrization (Section 3.5.1) and then bound the number of subsets (Section 3.5.2).

3.5.1 The VC Analysis: Symmetrization

We start with a couple of definitions.

Definition 3.6 (Dichotomies)**.** *Let $x_1, \ldots, x_n \in \mathcal{X}$. The set of* dichotomies *(the labeling patterns) generated by $\mathcal{H}$ on $x_1, \ldots, x_n$ is defined by*

$$\mathcal{H}(x_1, \ldots, x_n) = \{h(x_1), \ldots, h(x_n) : h \in \mathcal{H}\}.$$

Definition 3.7 (The Growth Function)**.** *The* growth function *of $\mathcal{H}$ is the maximal number of dichotomies it can generate on n points:*

$$m_{\mathcal{H}}(n) = \max_{x_1, \ldots, x_n} |\mathcal{H}(x_1, \ldots, x_n)|.$$

Put attention that $m_{\mathcal{H}}(n)$ is defined by the "worst-case" configuration of points $x_1, \ldots, x_n$, for which $|\mathcal{H}(x_1, \ldots, x_n)|$ is maximized. Thus, for lower bounding $m_{\mathcal{H}}(n)$ (i.e., for showing that $m_{\mathcal{H}}(n) \geq v$ for some value v) we have to find a configuration of points $x_1, \ldots, x_n$ for which $|\mathcal{H}(x_1, \ldots, x_n)| \geq v$ or, at least, prove that such configuration exists. For upper bounding $m_{\mathcal{H}}(n)$ (i.e., for showing that $m_{\mathcal{H}}(n) \leq v$) we have to show that for any possible configuration of points $x_1, \ldots, x_n$ we have $|\mathcal{H}(x_1, \ldots, x_n)| \leq v$. In other words, coming up with an example of a particular configuration $x_1, \ldots, x_n$ for which $|\mathcal{H}(x_1, \ldots, x_n)| \leq v$ is insufficient for proving that $m_{\mathcal{H}}(n) \leq v$, because there may potentially be an alternative configuration of points achieving a larger number of labeling configurations. To be concrete, the illustration in Figure 3.5b shows that for the hypothesis space $\mathcal{H}$ of homogeneous linear separators in $\mathbb{R}^2$ we have $m_{\mathcal{H}}(3) \geq 6$, but it does not show that $m_{\mathcal{H}}(3) \leq 6$. If we want to prove that $m_{\mathcal{H}}(3) \leq 6$ we have to show that no configuration of 3 sample points can differentiate between more than 6 distinct subsets of the hypothesis space. More generally, if we want to show that $m_{\mathcal{H}}(n) = v$ we have to show that $m_{\mathcal{H}}(n) \geq v$ and $m_{\mathcal{H}}(n) \leq v$. I.e., the only way to show equality is by proving a lower and an upper bound.

The following theorem uses the growth function to bound the distance between empirical and expected loss for all $h \in \mathcal{H}$.

Theorem 3.8. *Assume that ℓ is bounded in the $[0, 1]$ interval. Then for any $\delta \in (0, 1)$*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{8 \ln \frac{2 m_{\mathcal{H}}(2n)}{\delta}}{n}}\right) \leq \delta.$$

The result is useful when $m_{\mathcal{H}}(2n) \ll e^n$. In Section 3.5.2 we discuss when we can and cannot expect to have it, but for now we concentrate on the proof of the theorem.

The proof of the theorem is based on three ingredients. First we introduce a "ghost sample" S', which is an imaginary sample of the same size as S (i.e., of size n). We do not need to have this sample

[1] Homogeneous linear separators are linear separators passing through the origin.

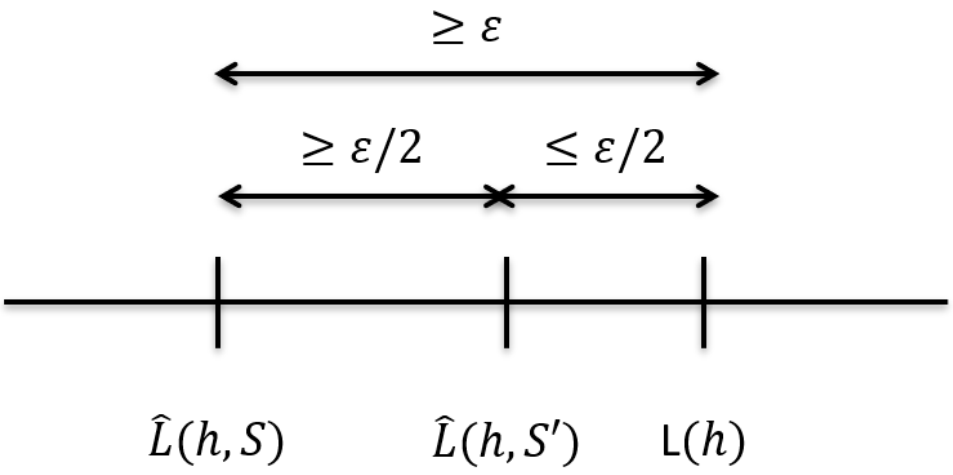


Figure 3.6: **Illustration for Step 2 of the proof of Theorem 3.8.**

at hand, but we ask what would have happened if we had such sample. Then we apply symmetrization: we show that the probability that for any h the empirical loss $\hat{L}(h,S)$ is far from $L(h)$ by more than ε is bounded by twice the probability that $\hat{L}(h,S)$ is far from $\hat{L}(h,S')$ by more than $\varepsilon/2$. This allows us to consider the behavior of $\mathcal{H}$ on the two samples, S and S', instead of studying it over all $\mathcal{X}$ (because the definition of $L(h)$ involves all $\mathcal{X}$, whereas the definition of $\hat{L}(h,S')$ involves only S'). In the third step we project $\mathcal{H}$ onto the two samples, S and S'. Even though $\mathcal{H}$ is uncountably infinite, when we look at it through the prism of $S \cup S'$ we can only observe a finite number of distinct behaviors. More precisely, the number of different ways $\mathcal{H}$ can label $S \cup S'$ is at most $m_{\mathcal{H}}(2n)$. We show that the probability that for any of the possible ways to label $S \cup S'$ the empirical losses $\hat{L}(h,S)$ and $\hat{L}(h,S')$ diverge by more than $\varepsilon/2$ decreases exponentially with n.

Now we do this formally.

Step 1 We introduce a ghost sample $S' = \{(X'_1, Y'_1), \ldots, (X'_n, Y'_n)\}$ of size n.

Step 2 [Symmetrization] We prove the following result.

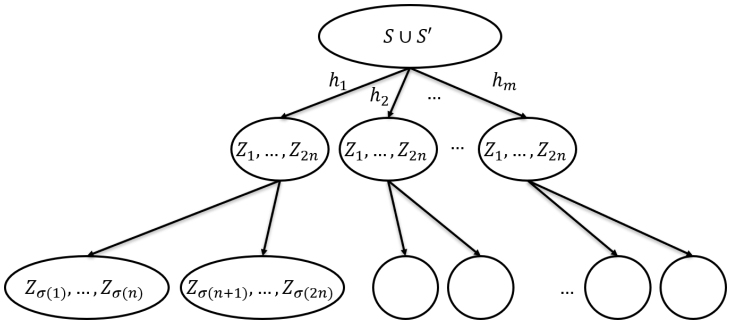
Lemma 3.9. *Assuming that $e^{-n\varepsilon^2/2} \leq \frac{1}{2}$ we have*

$$\mathbb{P}\Big(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\Big) \leq 2\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\Big). \tag{3.5}$$

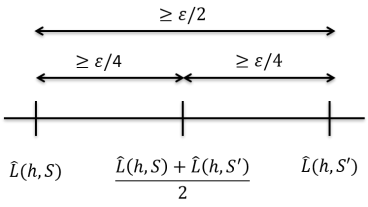
The illustration in Figure 3.6 should be helpful for understanding the proof. The distance $L(h) - \hat{L}(h,S)$ can be expressed as $L(h) - \hat{L}(h,S) = (L(h) - \hat{L}(h,S')) + (\hat{L}(h,S') - \hat{L}(h,S))$. We remind that in general empirical losses are likely to be close to their expected values. More explicitly, under the mild assumption that $e^{-n\varepsilon^2/2} \leq 1/2$ we have that $L(h) - \hat{L}(h,S') \leq \varepsilon/2$ with probability greater than $1/2$. If $L(h) - \hat{L}(h,S) \geq \varepsilon$ and $L(h) - \hat{L}(h,S') \leq \varepsilon/2$ we must have $\hat{L}(h,S') - \hat{L}(h,S) \geq \varepsilon/2$ (see the illustration). The proof is based on a careful exploitation of this observation.

Proof of Lemma 3.9. We start from the right hand side of (3.5).

$$\begin{aligned}
&\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\Big) \\
&\quad \geq \mathbb{P}\Big(\Big(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\Big) \text{ AND } \Big(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\Big)\Big) \\
&\quad = \mathbb{P}\Big(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\Big)\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\Big|\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\Big).
\end{aligned} \tag{3.6}$$



(a) Illustration of the split.



(b) Illustration of the distances.

Figure 3.7: **Illustration of the split of $S \cup S'$ into S and S'.** On the left: First we sample the joint sample $S \cup S'$. Then each hypothesis h_j produces a "big bag" of losses $\{Z_1, \ldots, Z_{2n}\}$, where $Z_i = \ell(h_j(X_i), Y_i)$. Even though $\mathcal{H}$ is uncountably infinite, the number of different ways to label $S \cup S'$ is at most $m_{\mathcal{H}}(2n)$ by the definition of the growth function and thus the number of different "big bags" of losses is at most $m_{\mathcal{H}}(2n)$ (in the illustration we have $m \leq m_{\mathcal{H}}(2n)$). Finally, we split $S \cup S'$ into S and S', which corresponds to splitting the "big bags" of $2n$ losses into pairs of "small bags" of n losses, corresponding to $\hat{L}(h_j, S)$ and $\hat{L}(h_j, S')$. On the right: we illustrate the distances between the average losses in a pair of "small bags" and the corresponding "big bag", which is the average of the two "small bags".

The inequality follows by the fact that for any two events A and B we have $\mathbb{P}(A) \geq \mathbb{P}(A \text{ AND } B)$ and the equality by $\mathbb{P}(A \text{ AND } B) = \mathbb{P}(B)\mathbb{P}(A|B)$. The first term in (3.6) is the term we want and we need to lower bound the second term. We let h^* be any h for which, by conditioning, we have $L(h^*) - \hat{L}(h^*, S) \geq \varepsilon$. With high probability we have that $\hat{L}(h^*, S')$ is close to $L(h^*)$ up to $\varepsilon/2$. And since we are given that $\hat{L}(h, S)$ is far from $L(h^*)$ by more than ε it must also be far from $\hat{L}(h^*, S')$ by more than $\varepsilon/2$ with high probability, see the illustration in Figure 3.6. Formally, we have:

$$
\begin{aligned}
&\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2} \Big| \exists h \in \mathcal{H} : L(h) - \hat{L}(h, S) \geq \varepsilon\Big) \\
&\qquad \geq \mathbb{P}\Big(\hat{L}(h^*, S') - \hat{L}(h^*, S) \geq \frac{\varepsilon}{2} \Big| L(h^*) - \hat{L}(h^*, S) \geq \varepsilon\Big) && (3.7)\\
&\qquad \geq \mathbb{P}\Big(L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2} \Big| L(h^*) - \hat{L}(h^*, S) \geq \varepsilon\Big) && (3.8)\\
&\qquad = \mathbb{P}\Big(L(h^*) - \hat{L}(h^*, S') \leq \frac{\varepsilon}{2}\Big) && (3.9)\\
&\qquad \geq 1 - \mathbb{P}\Big(L(h^*) - \hat{L}(h^*, S') \geq \frac{\varepsilon}{2}\Big) \\
&\qquad \geq 1 - e^{-2n(\varepsilon/2)^2} && (3.10)\\
&\qquad \geq \frac{1}{2}. && (3.11)
\end{aligned}
$$

Explanation of the steps: in (3.7) the event on the left hand side includes the event on the right hand side; in (3.8) we have $\hat{L}(h, S') - \hat{L}(h, S) = \Big(L(h) - \hat{L}(h, S)\Big) - \Big(L(h) - \hat{L}(h, S')\Big)$ and since we are given that $L(h) - \hat{L}(h, S) \geq \varepsilon$ the event $\hat{L}(h, S') - \hat{L}(h, S) \geq \varepsilon/2$ follows from $L(h) - \hat{L}(h, S') \leq \varepsilon/2$, see Figure 3.6; in (3.9) we can remove the conditioning on S, because the event of interest concerns S', which is independent of S; (3.10) follows by Hoeffding's inequality; and (3.11) follows by the lemma's assumption on $e^{-n\varepsilon^2/2}$.

By plugging the result back into (3.6) and multiplying by 2 we obtain the statement of the lemma. □

Step 3 [Projection] Now we focus on $\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2}\Big)$, which concerns the behavior of $\mathcal{H}$ on two finite samples, S and S'. There are two possible ways to sample S and S'. The first is to sample S and then S'. An alternative way is to sample a joint sample $S_{2n} = S \cup S'$ and then split

it into S and S' by randomly assigning half of the samples into S and half into S'. The two procedures are equivalent and lead to the same distribution over S and S'. We focus on the second procedure. Its advantage is that once we have sampled $S \cup S'$ the number of ways to label it with hypotheses from $\mathcal{H}$ is finite, even though $\mathcal{H}$ is uncountably infinite. This way we turn an uncountably infinite problem into a finite problem. The number of different sequences of losses on $S \cup S'$ is at most the number of different ways to label it, which is at most the growth function $m_{\mathcal{H}}(2n)$ by definition. The probability of having $\hat{L}(h, S') - \hat{L}(h, S) \geq \varepsilon/2$ for a fixed h reduces to the probability of splitting a sequence of $2n$ losses into n and n losses and having more than $\varepsilon/2$ difference between the average of the two. The latter reduces to the problem of sampling n losses without replacement from a bag of $2n$ losses and obtaining an average which deviates from the bag's average by more than $\varepsilon/4$, see Figure 3.7. This probability can be bounded by Hoeffding's inequality for sampling without replacement and decreases as $e^{-n\varepsilon^2/8}$. Putting this together we obtain the following result.

Lemma 3.10.

$$\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2}\Big) \leq m_{\mathcal{H}}(2n) e^{-n\varepsilon^2/8}.$$

As you may guess, $m_{\mathcal{H}}(2n)$ comes from a union bound over the number of possible sequences of losses we may obtain with hypotheses from $\mathcal{H}$ on $S \cup S'$. We now prove the lemma formally.

Proof of Lemma 3.10.

$$\begin{aligned}
\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2}\Big) &= \sum_{S \cup S'} \mathbb{P}(S \cup S') \mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2} \Big| S \cup S'\Big) \\
&\leq \sup_{S \cup S'} \mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2} \Big| S \cup S'\Big).
\end{aligned}$$

Put attention that the conditional probabilities are with respect to the splitting of $S \cup S'$ into S and S'.

Let $\mathcal{Z}(S \cup S') = \{Z_1, \ldots, Z_{2n} : Z_i = \ell(h(X_i), Y_i), h \in \mathcal{H}\}$ be the set of all possible sequences of losses that can be obtained by applying $h \in \mathcal{H}$ to $S \cup S'$. Since there are at most $m_{\mathcal{H}}(2n)$ distinct ways to label $S \cup S'$ we have $|\mathcal{Z}(S \cup S')| \leq m_{\mathcal{H}}(2n)$. Let $\sigma : \{1, \ldots, 2n\} \to \{1, \ldots, 2n\}$ denote a permutation of indexes. We have

$$\begin{aligned}
&\sup_{S \cup S'} \mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h, S') - \hat{L}(h, S) \geq \frac{\varepsilon}{2} \Big| S \cup S'\Big) \\
&\quad = \sup_{S \cup S'} \mathbb{P}\left(\exists \{Z_1, \ldots, Z_{2n}\} \in \mathcal{Z}(S \cup S') : \frac{1}{n} \sum_{i=1}^{n} Z_{\sigma(i)} - \frac{1}{n} \sum_{i=n+1}^{2n} Z_{\sigma(i)} \geq \frac{\varepsilon}{2}\right) && (3.12) \\
&\quad \leq \sup_{S \cup S'} \sum_{\{Z_1, \ldots, Z_{2n}\} \in \mathcal{Z}(S \cup S')} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Z_{\sigma(i)} - \frac{1}{n} \sum_{i=n+1}^{2n} Z_{\sigma(i)} \geq \frac{\varepsilon}{2}\right) && (3.13) \\
&\quad = \sup_{S \cup S'} \sum_{\{Z_1, \ldots, Z_{2n}\} \in \mathcal{Z}(S \cup S')} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} Z_{\sigma(i)} - \frac{1}{2n} \sum_{i=1}^{2n} Z_i \geq \frac{\varepsilon}{4}\right) && (3.14) \\
&\quad \leq \sup_{S \cup S'} \sum_{\{Z_1, \ldots, Z_{2n}\} \in \mathcal{Z}(S \cup S')} e^{-n\varepsilon^2/8} && (3.15) \\
&\quad \leq \sup_{S \cup S'} m_{\mathcal{H}}(2n) e^{-n\varepsilon^2/8} && (3.16) \\
&\quad = m_{\mathcal{H}}(2n) e^{-n\varepsilon^2/8},
\end{aligned}$$

where (3.12) follows by the fact that $\mathcal{Z}(S \cup S')$ is the set of all possible losses on $S \cup S'$ and in the step of splitting $S \cup S'$ into S and S' and computing $\hat{L}(h, S')$ and $\hat{L}(h, S)$ we are splitting a "big bag" of $2n$ losses into two "small bags" of n and n; all that is left from $\mathcal{H}$ in the splitting process is $\mathcal{Z}(S \cup S')$; the probability in (3.12) is over the split of $S \cup S'$ into S and S', which is expressed by taking the first n elements of a random permutation σ of indexes into S' and the last n elements into S and the probability is over σ; in (3.13) we apply the union bound; for (3.14) see the illustration in Figure 3.7b; in (3.15) we apply Hoeffding's inequality for sampling without replacement (Theorem 2.22) to the process of randomly sampling n losses out of $2n$ and observing $\varepsilon/4$ deviation from the average; in (3.16) we apply the bound on $|\mathcal{Z}(S \cup S')|$. □

Step 4 [Putting Everything Together] All that is left for the proof of Theorem 3.8 is to put Lemmas 3.9 and 3.10 together.

Proof of Theorem 3.8. Assuming that $e^{-n\varepsilon^2/2} \leq 1/2$ we have by Lemmas 3.9 and 3.10:

$$\begin{aligned}\mathbb{P}\Big(\exists h \in \mathcal{H} : L(h) - \hat{L}(h,S) \geq \varepsilon\Big) &\leq 2\mathbb{P}\Big(\exists h \in \mathcal{H} : \hat{L}(h,S') - \hat{L}(h,S) \geq \frac{\varepsilon}{2}\Big) \\ &\leq 2m_{\mathcal{H}}(2n)e^{-n\varepsilon^2/8}.\end{aligned}$$

Note that if $e^{-n\varepsilon^2/2} > 1/2$ then $2m_{\mathcal{H}}(2n)e^{-n\varepsilon^2/8} > 1$ and the inequality is satisfied trivially (because probabilities are always upper bounded by 1).

By denoting the right hand side of the inequality by δ and solving for ε we obtain the result. □

3.5.2 Bounding the Growth Function: The VC-dimension

In Theorem 3.8 we relate the distance between the expected and empirical losses to the growth function of $\mathcal{H}$. Our next goal is to bound the growth function. In order to do so we introduce the concept of shattering and the VC dimension.

Definition 3.11. *A set of points $x_1, \ldots, x_n$ is* shattered *by $\mathcal{H}$ if functions from $\mathcal{H}$ can produce all possible binary labellings of $x_1, \ldots, x_n$ or, in other words, if*

$$\|\mathcal{H}(x_1, \ldots, x_n)\| = 2^n.$$

For example, the set of homogeneous linear separators in $\mathbb{R}^2$ shatters the two points in Figure 3.5a, but it does not shatter the three points in Figure 3.5b. Note that if two points lie on one line passing through the origin, they are not shattered by the set of homogeneous linear separators, because they always get the same label. Thus, we may have two sets of points of the same size, where one is shattered and the other is not.

Definition 3.12. *The* Vapnik-Chervonenkis *(*VC*)* dimension *of $\mathcal{H}$, denoted by $d_{\mathrm{VC}}(\mathcal{H})$ is the maximal number of points that can be shattered by $\mathcal{H}$. In other words,*

$$d_{\mathrm{VC}}(\mathcal{H}) = \max\{n | m_{\mathcal{H}}(n) = 2^n\}.$$

If $m_{\mathcal{H}}(n) = 2^n$ for all n, then $d_{\mathrm{VC}}(\mathcal{H}) = \infty$.

Similar to the growth function, if we want to show that $d_{\mathrm{VC}}(\mathcal{H}) = d$ we have to show that $d_{\mathrm{VC}}(\mathcal{H}) \geq d$ and $d_{\mathrm{VC}}(\mathcal{H}) \leq d$. For example, the illustration in Figure 3.5a provides a configuration of points that are shuttered by homogeneous separating hyperplanes in $\mathbb{R}^2$ and thus shows that the VC-dimension of homogeneous separating hyperplanes in $\mathbb{R}^2$ is at least 2. However, the illustration in Figure 3.5b *does not* demonstrate that the VC-dimension of homogeneous separating hyperplanes in R^2 is smaller than 3. If we want to show that the VC-dimension of homogeneous separating hyperplanes in $\mathbb{R}^2$ is smaller than 3 we have to prove that no configuration of 3 points can be shattered. It is not sufficient to show that one particular configuration of points cannot be shattered. In the same spirit, two points lying on the same line passing through the origin cannot be shattered by homogeneous linear separators, but this does not tell anything about the VC-dimension, because we have another configuration of two points in Figure 3.5a that can be shattered. It is possible to show that the VC-dimension of homogeneous separating hyperplanes in $\mathbb{R}^d$ is d and the VC-dimension of general separating hyperplanes in $\mathbb{R}^d$ (not necessarily passing through the origin) is $d + 1$, see Abu-Mostafa et al. (2012, Exercise 2.4).

The next theorem bounds the growth function in terms of the VC-dimension.

Theorem 3.13 (Sauer's Lemma)**.**

$$m_{\mathcal{H}}(n) \leq \sum_{i=0}^{d_{\mathrm{VC}}(\mathcal{H})} \binom{n}{i}. \tag{3.17}$$

We remind that the binomial coefficient $\binom{n}{k}$ counts the number of ways to pick k elements out of n and that for $n < k$ it is defined as $\binom{n}{k} = 0$. Thus, equation (3.17) is well-defined even when $n < d_{\mathrm{VC}}(\mathcal{H})$. We also remind that $\sum_{i=0}^{n} \binom{n}{i} = 2^n$, where 2^n is the number of all possible subsets of n elements, which is equal to the sum over i going from 0 to n to select i elements out of n. For $n \leq d_{\mathrm{VC}}(\mathcal{H})$ we have $m_{\mathcal{H}}(n) = 2^n$ and the inequality is satisfied trivially.

The proof of Theorem 3.13 slightly reminds the combinatorial proof of the binomial identity

$$\binom{n}{k} = \binom{n-1}{k} + \binom{n-1}{k-1}.$$

One way to count the number of ways to select k elements out of n on the right hand side is to take one element aside. If that element is selected, then we have $\binom{n-1}{k-1}$ possibilities to select $k-1$ additional elements out of the remaining $n-1$. If the element is not selected, then we have $\binom{n-1}{k}$ possibilities to select all k elements out of remaining $n-1$. The sets including the first element are disjoint from the sets excluding it, leading to the identity above.

We need one more definition for the proof of Theorem 3.13.

Definition 3.14. *Let $B(n, d)$ be the maximal number of possible ways to label n points, so that no $d+1$ points are shattered.*

By the definition, we have $m_{\mathcal{H}}(n) \leq B(n, d_{\mathrm{VC}}(\mathcal{H}))$.

Proof of Theorem 3.13. We prove by induction that

$$B(n, d) \leq \sum_{i=0}^{d} \binom{n}{i}. \tag{3.18}$$

For the induction base we have $B(n, 0) = 1 = \binom{n}{0}$: if no points are shattered there is just one way to label the points. If there would be more than one way, they would differ in at least one point and that point would be shattered. By the definition of binomial coefficients, which says that for $k > n$ we have $\binom{n}{k} = 0$, we also know that for $n < d$ we have $B(n, d) = B(n, n)$. In particular, $B(0, d) = B(0, 0) = 1$.

Now we proceed with induction on d and for each d we do an induction on n. We show that

$$B(n, d) \leq B(n-1, d) + B(n-1, d-1).$$

Let $\mathcal{S}$ be a maximal set of dichotomies (labeling patterns) on n points $x_1, \ldots, x_n$. We take one point aside, x_n, and split $\mathcal{S}$ into three disjoint subsets: $\mathcal{S} = \mathcal{S}^* \cup \mathcal{S}^+ \cup \mathcal{S}^-$. The set $\mathcal{S}^*$ contains dichotomies on n points that appear with just one sign on x_n, either positive or negative. The sets $\mathcal{S}^+$ and $\mathcal{S}^-$ contain all dichotomies that appear with both positive and negative sign on x_n, where the positive ones are collected in $\mathcal{S}^+$ and the negative ones are collected in $\mathcal{S}^-$. Thus, the sets $\mathcal{S}^+$ and $\mathcal{S}^-$ are identical except in their labeling of x_n, where in $\mathcal{S}^+$ it is always labeled as $+$ and in $\mathcal{S}^-$ always as $-$. By contradiction, the number of points $x_1, \ldots, x_{n-1}$ that are shattered by $\mathcal{S}^-$ cannot be larger than $d-1$, because otherwise the number of points that are shattered by $\mathcal{S}$, which includes $\mathcal{S}^+$ and $\mathcal{S}^-$, would be larger than d, since we can use $\mathcal{S}^+$ and $\mathcal{S}^-$ to add x_n to the set of shattered points. Therefore, $|\mathcal{S}^-| \leq B(n-1, d-1)$. At the same time, the number of points $x_1, \ldots, x_{n-1}$ that are shattered by $\mathcal{S}^* \cup \mathcal{S}^+$ cannot be larger than d, because the total number of points shattered by $\mathcal{S}$ is at most d. Thus, we have $|\mathcal{S}^* \cup \mathcal{S}^+| \leq B(n-1, d)$. And overall

$$B(n, d) = |\mathcal{S}| = |\mathcal{S}^* \cup \mathcal{S}^+| + |\mathcal{S}^-| \leq B(n-1, d) + B(n-1, d-1),$$

as desired. By the induction assumption equation (3.18) is satisfied for $B(n-1, d)$ and $B(n-1, d-1)$, and we have

$$\begin{aligned} B(n, d) &\leq \sum_{i=0}^{d} \binom{n-1}{i} + \sum_{i=0}^{d-1} \binom{n-1}{i} \\ &= 1 + \sum_{i=0}^{d-1} \left(\binom{n-1}{i+1} + \binom{n-1}{i} \right) \\ &= \sum_{i=0}^{d} \binom{n}{i}, \end{aligned}$$

as desired. Finally, as we have already observed, $m_{\mathcal{H}}(n) \leq B(n, d_{\text{VC}}(\mathcal{H}))$, completing the proof. □

The following lemma provides a more explicit bound on the growth function.

Lemma 3.15.

$$\sum_{i=0}^{d} \binom{n}{i} \leq n^d + 1.$$

The proof is based on induction and left as an exercise.

By plugging the results of Theorem 3.13 and Lemma 3.15 into Theorem 3.8 we obtain the VC generalization bound.

Theorem 3.16 (VC generalization bound)**.** *Let $\mathcal{H}$ be a hypotheses class with VC-dimension $d_{\text{VC}}(\mathcal{H}) = d_{\text{VC}}$. Then:*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \hat{L}(h, S) + \sqrt{\frac{8 \ln\left(2\left((2n)^{d_{\text{VC}}} + 1\right)/\delta\right)}{n}} \right) \leq \delta.$$

For example, the VC-dimension of linear separators in $\mathbb{R}^d$ is $d + 1$ and theorem 3.16 provides generalization guarantees for learning with linear separators in finite-dimensional spaces, as long as the dimension of the space d is small in relation to the number of points n.

3.6 VC Analysis of SVMs

Kernel Support Vector Machines (SVMs) can map the data into high and potentially infinite-dimensional spaces. For example, Radial Basis Function (RBF) kernels map the data into an infinite-dimensional space. In the following we provide a more refined analysis of generalization in learning with linear separators in high-dimensional spaces. The analysis is based on the notion of *separation with a margin*. We use the following definitions.

Definition 3.17 (Fat Shattering)**.** *Let $\mathcal{H}_\gamma = \{(\mathbf{w}, b) : \|\mathbf{w}\| \leq 1/\gamma\}$ be the space of hyperplanes described by $\mathbf{w}$ and b, where $\mathbf{w}$ is a vector in $\mathbb{R}^d$ (with potentially infinite dimension d) with $\|\mathbf{w}\| \leq 1/\gamma$ and $b \in \mathbb{R}$. We say that a set of points $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ is* fat-shattered *by $\mathcal{H}_\gamma$ if for any set of labels $\{y_1, \ldots, y_n\} \in \{\pm 1\}^n$ we have a hyperplane $(\mathbf{w}, b) \in \mathcal{H}_\gamma$ that satisfies $y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1$ for all $i \in \{1, \ldots, n\}$.*

Note that when $y = \text{sign}\left(\langle \mathbf{w}, \mathbf{x} \rangle + b\right)$ the distance of a point $\mathbf{x}$ to a hyperplane h defined by $(\mathbf{w}, b)$ is given by $\text{dist}(h, \mathbf{x}) = \frac{y(\langle \mathbf{w}, \mathbf{x} \rangle + b)}{\|\mathbf{w}\|}$ (Abu-Mostafa et al., 2015, Page 5, Chapter 8) and for $h = (\mathbf{w}, b) \in \mathcal{H}_\gamma$ and $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ fat-shattered by $\mathcal{H}_\gamma$ we obtain $\text{dist}(h, \mathbf{x}_i) \geq \gamma$ for all $i \in \{1, \ldots, n\}$. It means that any possible labeling of $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ can be achieved with margin at least γ.

Definition 3.18 (Fat Shattering Dimension)**.** *We say that* fat shattering dimension $d_{\text{FAT}}(\mathcal{H}_\gamma) = d$ *if d is the maximal number of points that can be fat shattered by $\mathcal{H}_\gamma$. (I.e., there exist n points that can be fat shattered by $\mathcal{H}_\gamma$ and no $d + 1$ points can be fat shattered by $\mathcal{H}_\gamma$.)*

Note that $d_{\text{FAT}}(\mathcal{H}_\gamma) \leq d_{\text{VC}}(\mathcal{H}_\gamma) \leq d + 1$, where d is the dimension of $\mathbf{w}$. (If we can shatter n points with margin γ we can also shatter them without the margin.)

The following theorem bounds the fat shattering dimension of $\mathcal{H}_\gamma$, see Abu-Mostafa et al. (2015) for a proof.

Theorem 3.19 ((Abu-Mostafa et al., 2015, Theorem 8.5))**.** *Assume that the input space $\mathcal{X}$ is a ball of radius R in $\mathbb{R}^d$ (i.e., $\|x\| \leq R$ for all $x \in \mathcal{X}$), where d may potentially be infinite. Then:*

$$d_{\text{FAT}}(\mathcal{H}_\gamma) \leq \lceil R^2/\gamma^2 \rceil + 1,$$

where $\lceil R^2/\gamma^2 \rceil$ is the smallest integer that is greater or equal to R^2/γ^2.

The important point is that the bound on fat shattering dimension is independent of the dimension of the space $\mathbb{R}^d$ that $\mathbf{w}$ comes from.

We define fat losses that count as error everything that falls too close to the separating hyperplane or on the wrong side of it.

Definition 3.20 (Fat Losses). *For $h = (\mathbf{w}, b)$ we define the fat losses*

$$\ell_{\text{FAT}}(h(\mathbf{x}), y) = \begin{cases} 0, & \text{if } y_i(\langle \mathbf{w}, \mathbf{x}_i \rangle + b) \geq 1 \\ 1, & \text{otherwise,} \end{cases}$$
$$L_{\text{FAT}}(h) = \mathbb{E}\left[\ell_{\text{FAT}}(h(\mathbf{X}), Y)\right],$$
$$\hat{L}_{\text{FAT}}(h, S) = \frac{1}{n} \sum_{i=1}^{n} \ell_{\text{FAT}}(h(\mathbf{X}_i), Y_i).$$

In relation to the fat losses the fat shattering dimension acts in the same way as the VC-dimension in relation to the zero-one loss. In particular, we have the following result that relates $L_{\text{FAT}}(h)$ to $\hat{L}_{\text{FAT}}(h, S)$ via $d_{\text{FAT}}(\mathcal{H}_\gamma)$ (the proof is left as an exercise).

Theorem 3.21.

$$\mathbb{P}\left(\exists h \in \mathcal{H}_\gamma : L_{\text{FAT}}(h) \geq \hat{L}_{\text{FAT}}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{d_{\text{FAT}}(\mathcal{H}_\gamma)} + 1 \right) / \delta \right)}{n}} \right) \leq \delta.$$

Now we are ready to analyze generalization in learning with fat linear separation. For the analysis we make a simplifying assumption that the data are contained within a ball of radius $R = 1$. The analysis for general R is left as an exercise. Note that R refers to the radius of the ball *after* potential transformation of the data through a feature mapping / kernel function. For example, the RBF kernel maps the data into an infinite dimensional space and we consider the radius of the ball containing the transformed data in the infinite dimensional space.

Theorem 3.22. *Assume that the input space $\mathcal{X}$ is a ball of radius $R = 1$ in $\mathbb{R}^d$, where d is potentially infinite. Let $\mathcal{H}$ be the space of linear separators $h = (\mathbf{w}, b)$. Then*

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L_{\text{FAT}}(h) \geq \hat{L}_{\text{FAT}}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+\lceil \|\mathbf{w}\|^2 \rceil} + 1 \right) (1 + \lceil \|\mathbf{w}\|^2 \rceil) \lceil \|\mathbf{w}\|^2 \rceil / \delta \right)}{n}} \right) \leq \delta.$$

Observe that $L(h) \leq L_{\text{FAT}}(h)$ and, therefore, the theorem provides a generalization bound for $L(h)$. (If we count correct classifications within the margin as errors we only increase the loss.)

Proof. The proof is based on combination of VC and Occam's razor bounding techniques, see the illustration in Figure 3.8. We start by noting that Theorem 3.19 is interesting when $\lceil R^2/\gamma^2 \rceil < d + 1$, because as we have already noted $d_{\text{FAT}}(\mathcal{H}_\gamma) \leq d_{\text{VC}}(\mathcal{H}_\gamma) \leq d + 1$. We slice the hypotheses space $\mathcal{H}$ into a nested sequence of subspaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \cdots \subset \mathcal{H}_{d-1} \subset \mathcal{H}_d = \mathcal{H}$, where for all $i < d$ we define $\mathcal{H}_i$ to be the hypothesis space $\mathcal{H}_\gamma$ with $1/\gamma^2 = i$. In other words, $\mathcal{H}_i = \mathcal{H}_{\left\{\gamma = \frac{1}{\sqrt{i}}\right\}}$ (do not let the notation to confuse you, by $\mathcal{H}_i$ we denote the i-th hypothesis space in the nested sequence of hypothesis spaces and by $\mathcal{H}_\gamma$ we denote the hypothesis space with $\|\mathbf{w}\|$ upper bounded by $1/\gamma$). By Theorem 3.19 we have $d_{\text{FAT}}(\mathcal{H}_i) = i + 1$ and then by Theorem 3.21:

$$\mathbb{P}\left(\exists h \in \mathcal{H}_i : L_{\text{FAT}}(h) \geq \hat{L}_{\text{FAT}}(h, S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+i} + 1 \right) / \delta_i \right)}{n}} \right) \leq \delta_i.$$

We take $\delta_i = \frac{1}{i(i+1)} \delta$ and note that $\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \sum_{i=1}^{\infty} \left(\frac{1}{i} - \frac{1}{i+1} \right) = \left(1 - \frac{1}{2}\right) + \left(\frac{1}{2} - \frac{1}{3}\right) + \left(\frac{1}{3} - \frac{1}{4}\right) + \cdots = 1$. We also note that $\mathcal{H} = \bigcup_{i=1}^{d} (\mathcal{H}_i \setminus \mathcal{H}_{i-1})$, where $\mathcal{H}_0$ is defined as the empty set and $\mathcal{H}_i \setminus \mathcal{H}_{i-1}$ is the difference between sets $\mathcal{H}_i$ and $\mathcal{H}_{i-1}$ (everything that is in $\mathcal{H}_i$, but not in $\mathcal{H}_{i-1}$). Note that the sets

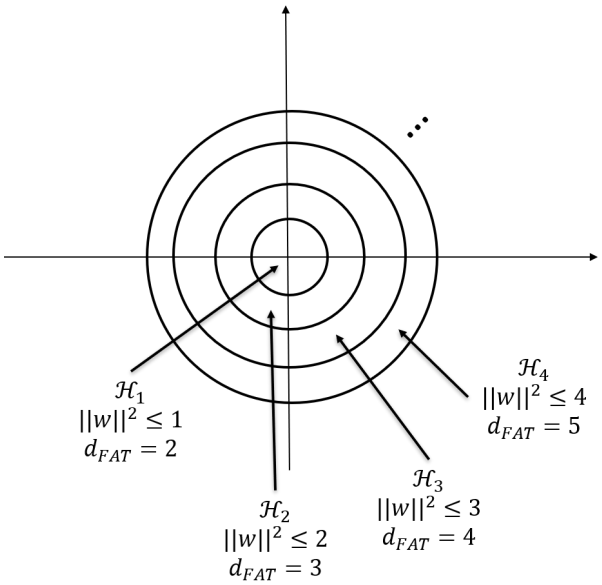


Figure 3.8: **Illustration for the proof of Theorem 3.22**

$\mathcal{H}_i \setminus \mathcal{H}_{i-1}$ and $\mathcal{H}_j \setminus \mathcal{H}_{j-1}$ are disjoint for $i \neq j$. Also note that δ_i is a distribution of our confidence budget δ among $\mathcal{H}_i \setminus \mathcal{H}_{i-1}$-s. Finally, note that if $h = (\mathbf{w}, b) \in \mathcal{H}_i \setminus \mathcal{H}_{i-1}$ then $\lceil \|\mathbf{w}\|^2 \rceil = i$. The remainder of the proof follows the same lines as the proof of Occam's razor bound:

$$
\begin{aligned}
&\mathbb{P}\left(\exists h \in \mathcal{H} : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+\lceil \|\mathbf{w}\|^2 \rceil} + 1 \right) (1 + \lceil \|\mathbf{w}\|^2 \rceil) \lceil \|\mathbf{w}\|^2 \rceil / \delta \right)}{n}} \right) \\
&\quad = \mathbb{P}\left(\exists h \in \bigcup_{i=1}^{d} \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+\lceil \|\mathbf{w}\|^2 \rceil} + 1 \right) (1 + \lceil \|\mathbf{w}\|^2 \rceil) \lceil \|\mathbf{w}\|^2 \rceil / \delta \right)}{n}} \right) \\
&\quad = \sum_{i=1}^{d} \mathbb{P}\left(\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+\lceil \|\mathbf{w}\|^2 \rceil} + 1 \right) (1 + \lceil \|\mathbf{w}\|^2 \rceil) \lceil \|\mathbf{w}\|^2 \rceil / \delta \right)}{n}} \right) \\
&\quad = \sum_{i=1}^{d} \mathbb{P}\left(\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+i} + 1 \right) (1+i)\, i / \delta \right)}{n}} \right) \\
&\quad = \sum_{i=1}^{d} \mathbb{P}\left(\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+i} + 1 \right) / \delta_i \right)}{n}} \right) \\
&\quad \leq \sum_{i=1}^{d} \mathbb{P}\left(\exists h \in \mathcal{H}_i : L_{\mathrm{FAT}}(h) \geq \hat{L}_{\mathrm{FAT}}(h,S) + \sqrt{\frac{8 \ln \left(2 \left((2n)^{1+i} + 1 \right) / \delta_i \right)}{n}} \right) \\
&\quad \leq \sum_{i=1}^{d} \delta_i = \sum_{i=1}^{d} \frac{1}{i(i+1)} \delta = \delta \sum_{i=1}^{d} \frac{1}{i(i+1)} \leq \delta \sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \delta.
\end{aligned}
$$

□

3.7 VC Lower Bound

In this section we show that when the VC-dimension is unbounded, it is impossible to bound the distance between $L(h)$ and $\hat{L}(h,S)$.

Theorem 3.23. *Let $\mathcal{H}$ be a hypothesis class with $d_{VC}(\mathcal{H}) = \infty$. Then for any n there exists a distribution over $\mathcal{X}$ and a class of target functions $\mathcal{F}$, such that*

$$\mathbb{E}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right] \geq 0.25,$$

where the expectation is over selection of a sample of size n and a target function.

Proof. Pick n. Since $d_{VC}(\mathcal{H}) = \infty$ we know that there exist $2n$ points that are shattered by $\mathcal{H}$. Let the sample space $\mathcal{X}_{2n} = \{x_1, \ldots, x_{2n}\}$ be these points and let $p(x)$ be uniform on $\mathcal{X}_{2n}$. Let $\mathcal{F}$ be the set of all possible functions from $\mathcal{X}_{2n}$ to $\{0,1\}$ and let $p(f)$ be uniform over $\mathcal{F}$. Let S be a sample of n points. Let $\{\mathcal{F}_k(S)\}_k$ be maximal subsets of $\mathcal{F}$, such that $\mathcal{F} = \bigcup_k \mathcal{F}_k(S)$ and any $f_i, f_j \in \mathcal{F}_k(S)$ agree on S. Note that since $\mathcal{X}_{2n}$ is shattered by $\mathcal{H}$, for any S, any $\mathcal{F}_k$, and any $f_i \in \mathcal{F}_k$ that was used to label S there exists $h^*(\mathcal{F}_k(S), S) \in \mathcal{H}$, such that for any $f_i \in \mathcal{F}_k(S)$ the empirical error $\hat{L}(h^*(f_i, S), S) = 0$. Let $p(k)$ and $p(i)$ be uniform. Then:

$$\begin{aligned}
\mathbb{E}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right] &= \mathbb{E}_{f\sim p(f)}\left[\mathbb{E}_{S\sim p(X)^n}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right]\Big| f\right] \\
&= \mathbb{E}_{S\sim p(X)^n}\left[\mathbb{E}_{f\sim p(f)}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right]\Big| S\right] \\
&= \mathbb{E}_{S\sim p(X)^n}\left[\mathbb{E}_{k\sim p(k)}\left[\mathbb{E}_{i\sim p(i)}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right]\Big|\mathcal{F}_k\right]\Big| S\right] \\
&\geq \mathbb{E}_{S\sim p(X)^n}\left[\mathbb{E}_{k\sim p(k)}\left[\mathbb{E}_{i\sim p(i)}\left[L(h^*(\mathcal{F}_k,S)) - \hat{L}(h^*(\mathcal{F}_k,S),S)\right]\Big|\mathcal{F}_k\right]\Big| S\right] \\
&= \mathbb{E}_{S\sim p(X)^n}\left[\mathbb{E}_{k\sim p(k)}\left[\mathbb{E}_{i\sim p(i)}\left[L(h^*(\mathcal{F}_k,S))\right]\Big|\mathcal{F}_k\right]\Big| S\right] \\
&= \mathbb{E}_{S\sim p(X)^n}\left[\mathbb{E}_{k\sim p(k)}\left[0.25\right]\Big| S\right] \\
&= 0.25.
\end{aligned}$$

□

Corollary 3.24. *Under the assumptions of Theorem 3.23, with probability at least 0.125,* $\sup_h(L(h) - \hat{L}(h,S)) \geq 0.125$. *Thus, it is impossible to have high-probability bounds on* $\sup_h(L(h) - \hat{L}(h,S))$ *that converge to zero as n goes to infinity.*

Proof. Note that $\sup_h(L(h) - \hat{L}(h,S)) \leq 1$, since ℓ is bounded in $[0,1]$. Assume by contradiction that $\mathbb{P}\Big(\sup_h \left(L(h) - \hat{L}(h,S)\right) \geq 0.125\Big) < 0.125$. Then

$$\mathbb{E}\left[\sup_h \left(L(h) - \hat{L}(h,S)\right)\right] \leq 0.125 \times 1 + (1 - 0.125) \times 0.125 < 2 \times 0.125 = 0.25,$$

which is in contradiction with Theorem 3.23. □

3.8 PAC-Bayesian Analysis

Occam's razor and VC analysis consider hard selection of a single hypothesis from a hypothesis class. In PAC-Bayesian analysis hard selection is replaced by a soft selection: instead of selecting a single hypothesis, it is allowed to select a distribution over the hypothesis space. When the distribution is a delta-distribution putting all the mass on a single hypothesis, hard selection is recovered and the outcome is identical to Occam's razor bound. However, the possibility of soft selection provides much more freedom and control over the approximation-estimation trade-off. PAC-Bayesian generalization bounds are based on *change of measure* inequality, which acts as replacement for the union bound. Change of measure

inequality has two important advantages over the union bound: (1) it is tighter (you will verify this in a home assignment) and (2) it can be applied to uncountably infinite hypothesis classes. Furthermore, soft selection allows application of gradient-descent type methods to optimize the distribution over $\mathcal{H}$, which in some cases leads to efficient algorithms for direct minimization of the PAC-Bayesian bounds. Soft selection is implemented by *randomized classifiers*, which are formally defined below.

Definition 3.25 (Randomized Classifier)**.** *Let ρ be a distribution over $\mathcal{H}$. A* randomized classifier *associated with ρ (and named ρ) acts according to the following scheme. At each prediction round it:*

1. *Picks $h \in \mathcal{H}$ according to $\rho(h)$*
2. *Observes x*
3. *Returns $h(x)$*

The expected loss of ρ is $\mathbb{E}_{h\sim\rho}[L(h)]$ and the empirical loss is $\mathbb{E}_{h\sim\rho}\left[\hat{L}(h,S)\right]$. Whenever it does not lead to confusion, we will shorten the notation to $\mathbb{E}_{\rho}[L(h)]$ and $\mathbb{E}_{\rho}\left[\hat{L}(h,S)\right]$.

There is a large number of different PAC-Bayesian inequalities. We start with the classical one due to Seeger (2002).

Theorem 3.26 (PAC-Bayes-kl inequality)**.** *For any "prior" distribution π over $\mathcal{H}$ that is independent of S, for all randomized classifiers (distributions) ρ simultaneously:*

$$\mathbb{P}\left(\mathrm{kl}\left(\mathbb{E}_{\rho}\left[\hat{L}(h,S)\right]\middle\|\mathbb{E}_{\rho}\left[L(h)\right]\right) \geq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}\right) \leq \delta. \tag{3.19}$$

The meaning of "prior" should be interpreted in exactly the same way as the "prior" in Occam's razor bound: it is any distribution over $\mathcal{H}$ that sums up to one and does not depend on the sample S. The prior is an auxiliary construction for deriving the bound and unlike in Bayesian learning there is no assumption that it reflects any real-world distribution over $\mathcal{H}$.

Before proceeding to the proof of the theorem we provide a discussion of its meaning. To get some intuition we apply Pinsker's relaxation of kl (inequality 2.17) that leads to a more digestible (although weaker) form of the bound: with probability greater than $1-\delta$ for all ρ simultaneously

$$\mathbb{E}_{\rho}[L(h)] \leq \mathbb{E}_{\rho}\left[\hat{L}(h,S)\right] + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{2n}}.$$

Note that when $\rho = \pi$ the KL term is zero and we recover generalization bound for a single hypothesis. Taking $\rho = \pi$ amounts to making no selection. If we start with a prior distribution π and continue with it without taking any information from the sample we get the usual Hoeffding's or kl inequality. In order to get more intuition about the bound we decompose the KL-divergence:

$$\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{\rho}\left[\ln\frac{\rho}{\pi}\right] = \underbrace{\mathbb{E}_{\rho}\left[\ln\frac{1}{\pi}\right]}_{\substack{\text{Average}\\ \text{complexity}}} - \underbrace{\mathrm{H}(\rho)}_{\text{Entropy}}.$$

If $\mathcal{H}$ is finite and π is uniform, then $\mathrm{H}(\rho) \geq 0$ and $\mathrm{KL}(\rho\|\pi) = \ln|\mathcal{H}| - \mathrm{H}(\rho) \leq \ln|\mathcal{H}|$ and we recover generalization bound for finite hypothesis sets with an improvement by $-\mathrm{H}(\rho)$. Recall that the entropy $\mathrm{H}(\rho)$ is zero when ρ is a delta-distribution and when ρ is uniform the entropy has its maximal value, which is $\ln|\mathcal{H}|$. Thus, $-\mathrm{H}(\rho)$ is an "award" for avoiding commitment to a single hypothesis.

Overall, the PAC-Bayesian inequality advocates for picking ρ that minimizes the trade-off between:

1. The empirical error $\hat{L}(h,S)$.
2. The complexity (description length, prior belief) $\ln\frac{1}{\pi(h)}$.
3. And has maximum entropy (it has "indifference" to h and h' when $\hat{L}(h,S) = \hat{L}(h',S)$ and $\pi(h) = \pi(h')$). Maximization of $\mathrm{H}(\rho)$ corresponds to avoidance of selection whenever it is not necessary. Reduced selection leads to improved estimation without impairing the approximation and provides a tighter generalization bound.

3.8.1 Relation and Differences with other Learning Approaches

PAC-Bayesian analysis has the following relation and differences with Bayesian learning and with VC analysis / Radamacher complexities.

Relation with Bayesian learning

1. Explicit way to incorporate prior information (via $\pi(h)$).

Difference with Bayesian learning

1. Explicit high-probability guarantee on the expected performance.
2. No belief in prior correctness (frequentist bound).
3. Explicit dependence on the loss function.
4. Different weighting of prior belief $\pi(h)$ vs. evidence $\hat{L}(h)$.
5. Holds for *any* distribution ρ (including the Bayes posterior).

Relation with VC analysis / Radamacher complexities

1. Explicit high-probability guarantee on the expected performance.
2. Explicit dependence on the loss function.

Difference with VC analysis / Radamacher complexities

1. Complexity is defined individually for each h via $\pi(h)$ (rather than "complexity of a hypothesis class").
2. Explicit way to incorporate prior knowledge.
3. The bound is defined for randomized classifiers ρ (not individual h); but workarounds exist in some cases.

In a sense, PAC-Bayesian analysis takes the best out of Bayesian learning and VC analysis and puts it together. And it also leads to efficient learning algorithms, since $\mathrm{KL}(\rho\|\pi)$ is convex in ρ and $\hat{L}(\rho, S)$ is linear in ρ.

3.8.2 A Proof of PAC-Bayes-kl Inequality

At the basis of most of PAC-Bayesian bounds lies the change of measure inequality, which acts as replacement of the union bound for uncountably infinite sets.

Theorem 3.27 (Change of measure inequality)**.** *For any measurable function $f(h)$ on $\mathcal{H}$ and any distributions ρ and π:*

$$\mathbb{E}_{h\sim\rho(h)}\left[f(h)\right] \leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{h\sim\pi(h)}\left[e^{f(h)}\right].$$

Proof.

$$\begin{aligned}
\mathbb{E}_{\rho(h)}\left[f(h)\right] &= \mathbb{E}_{\rho(h)}\left[\ln\left(\frac{\rho(h)}{\pi(h)} \times e^{f(h)} \times \frac{\pi(h)}{\rho(h)}\right)\right] \\
&= \mathrm{KL}(\rho\|\pi) + \mathbb{E}_{\rho(h)}\left[\ln\left(e^{f(h)} \times \frac{\pi(h)}{\rho(h)}\right)\right] \\
&\leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\rho(h)}\left[e^{f(h)} \times \frac{\pi(h)}{\rho(h)}\right] \\
&= \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_{\pi(h)}\left[e^{f(h)}\right],
\end{aligned}$$

where the inequality in the third step is justified by Jensen's inequality (Theorem B.30). Note that there is nothing probabilistic in the statement of the theorem - it is a deterministic result. □

In the next lemma we extend f to be a function of h and a sample S and apply a probabilistic argument to the last term of change-of-measure inequality. The lemma is the foundation for most PAC-Bayesian bounds.

Lemma 3.28 (PAC-Bayes lemma)**.** *For any measurable function $f : \mathcal{H} \times (\mathcal{X} \times \mathcal{Y})^n \to \mathbb{R}$ and any distribution π over $\mathcal{H}$ that is independent of the sample S, for all distributions ρ simultaneously*

$$\mathbb{P}\left(\mathbb{E}_{h\sim\rho}\left[f(h,S)\right] \geq \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_{h\sim\pi}\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \leq \delta,$$

where the probability is with respect to the draw of the sample S and $\mathbb{E}_S$ is the expectation with respect to the draw of S.

With the usual rewriting, we obtain that with probability at least $1-\delta$ over the draw of S, for all ρ simultaneously

$$\mathbb{E}_\rho\left[f(h,S)\right] \leq \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_{\pi}\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}.$$

We first present a slightly less formal, but more intuitive proof and then provide a formal one. By change of measure inequality we have

$$\begin{aligned}
\mathbb{E}_\rho\left[f(h,S)\right] &\leq \mathrm{KL}(\rho\|\pi) + \ln\mathbb{E}_\pi\left[e^{f(h,S)}\right] \\
&\underset{w.p.\geq 1-\delta}{\leq} \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_S\left[\mathbb{E}_\pi\left[e^{f(h,S)}\right]\right]}{\delta} \\
&= \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta},
\end{aligned}$$

where in the second line we apply Markov's inequality to the random variable $Z = \mathbb{E}_\pi\left[e^{f(h,S)}\right]$ (and the inequality holds with probability at least $1-\delta$) and in the last line we can exchange the order of expectations, because π is independent of S. The key observation is that the change-of-measure inequality relates all posterior distributions ρ to a single prior distribution π in a deterministic way and the probabilistic argument (Markov's inequality) is applied to a single random quantity $\mathbb{E}_\pi\left[e^{f(h,S)}\right]$. This way change-of-measure inequality replaces the union bound and it holds even when $\mathcal{H}$ is uncountably infinite.

Now we provide a formal proof.

Proof.

$$\begin{aligned}
\mathbb{P}\left(\mathbb{E}_\rho\left[f(h,S)\right] \geq \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_{\pi}\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) &\leq \mathbb{P}\left(\ln\mathbb{E}_\pi\left[e^{f(h,S)}\right] \geq \ln\frac{\mathbb{E}_{\pi}\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \quad (3.20)\\
&= \mathbb{P}\left(\mathbb{E}_\pi\left[e^{f(h,S)}\right] \geq \frac{\mathbb{E}_{\pi}\left[\mathbb{E}_S\left[e^{f(h,S)}\right]\right]}{\delta}\right) \\
&= \mathbb{P}\left(\mathbb{E}_\pi\left[e^{f(h,S)}\right] \geq \frac{\mathbb{E}_{S}\left[\mathbb{E}_\pi\left[e^{f(h,S)}\right]\right]}{\delta}\right) \quad (3.21)\\
&\leq \delta,
\end{aligned}$$

where (3.20) follows by change-of-measure inequality, in (3.21) we can exchange the order of expectations, because π is independent of S, and in the last step we apply Markov's inequality to the random variable $Z = \mathbb{E}_\pi\left[e^{f(h,S)}\right]$.

We repeat that the change-of-measure inequality relates all posterior distributions ρ to a single prior distribution π in a deterministic way and the probabilistic argument in the last step is applied to a single random variable $\mathbb{E}_\pi\left[e^{f(h,S)}\right]$. □

Different PAC-Bayesian inequalities are obtained by different choices of the function $f(h,S)$. A key consideration in the choice of $f(h,S)$ is the possibility to bound the moment generating function

$\mathbb{E}_S\left[e^{f(h,S)}\right]$. For example, we have done it for $f(h,S) = n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))$ in Lemma 2.14 and this is the choice of f in the proof of PAC-Bayes-kl inequality. Other choices of f are possible. For example, Hoeffding's Lemma 2.6 provides a bound on the moment generating function of $f(h,S) = \lambda\left(L(h) - \hat{L}(h,S)\right)$, which can be used to derive PAC-Bayes-Hoeffding inequality. We refer to Seldin et al. (2012) for more details.

The proof of PAC-Bayes-kl inequality relies on convexity of the kl-divergence. We cite the theorem and refer to Cover and Thomas (2006) for details.

Theorem 3.29 (Cover and Thomas, 2006, Theorem 2.7.2)**.** $\mathrm{KL}(p\|q)$ *is convex in the pair* (p,q)*; that is, if* (p_1,q_1) *and* (p_2,q_2) *are two pairs of probability mass functions, then*

$$\mathrm{KL}(\lambda p_1 + (1-\lambda)p_2\|\lambda q_1 + (1-\lambda)q_2) \le \lambda\,\mathrm{KL}(p_1\|q_1) + (1-\lambda)\,\mathrm{KL}(p_2\|q_2)$$

for all $0 \le \lambda \le 1$.

Corollary 3.30.

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) \le \mathbb{E}_\rho\left[\mathrm{kl}\left(\hat{L}(h,S)\middle\|L(h)\right)\right].$$

Finally, we are ready to prove Theorem 3.26.

Proof of Theorem 3.26. We provide an intuitive derivation and leave the formal one (as in the proof of Lemma 3.28) as an exercise.

We take $f(h,S) = n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))$. Then we have

$$\begin{aligned}
n\,\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) &\le \mathbb{E}_\rho\left[n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))\right] \\
&\underset{w.p.\ge 1-\delta}{\le} \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right]\right]}{\delta} \\
&\le \mathrm{KL}(\rho\|\pi) + \ln\frac{\mathbb{E}_\pi\left[n+1\right]}{\delta} \\
&= \mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta},
\end{aligned}$$

where the first inequality is by Corollary 3.30, the second inequality is by the PAC-Bayes Lemma (and it holds with probability at least $1-\delta$ over the draw of S), and the third inequality is by Lemma 2.14. □

3.8.3 Application to SVMs

In order to apply PAC-Bayesian bound to a given problem we have to design a prior distribution π and then bound the KL-divergence $\mathrm{KL}(\rho\|\pi)$ for the posterior distributions of interest. Sometimes we resort to a restricted class of ρ-s, for which we are able to bound $\mathrm{KL}(\rho\|\pi)$. You can see how this is done for SVMs in Langford (2005, Section 5.3).

3.8.4 Relaxation of PAC-Bayes-kl: PAC-Bayes-λ Inequality

Due to its implicit form, PAC-Bayes-kl inequality is not very convenient for optimization. One way around is to replace the bound with a linear trade-off $\beta n\mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \mathrm{KL}(\rho\|\pi)$. Since $\mathrm{KL}(\rho\|\pi)$ is convex in ρ and $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ is linear in ρ, for a fixed β the trade-off is convex in ρ and can be minimized. (We note that parametrization of ρ, for example the popular restriction of ρ to a Gaussian posterior (Langford, 2005), may easily break the convexity (Germain et al., 2009). We get back to this point in Section 3.8.6.) The value of β can then be tuned by cross-validation or substitution of $\rho(\beta)$ into the bound (the former usually works better).

Below we present a more rigorous approach. We prove the following relaxation of PAC-Bayes-kl inequality, which leads to a bound that can be optimized by alternating minimization.

Theorem 3.31 (PAC-Bayes-λ Inequality)**.** *For any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ and all $\lambda \in (0,2)$ and $\gamma > 0$ simultaneously:*

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)n}, \tag{3.22}$$

$$\mathbb{E}_\rho\left[L(h)\right] \geq \left(1-\frac{\gamma}{2}\right)\mathbb{E}_\rho[\hat{L}(h,S)] - \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\gamma n}. \tag{3.23}$$

At the moment we focus on the upper bound in equation (3.22). Note that the theorem holds for *all* values of $\lambda \in (0,2)$ simultaneously. Therefore, we can optimize the bound with respect to λ and pick the best one.

Proof. We prove the upper bound in equation (3.22). Proof of the lower bound (3.23) is analogous and left as an exercise. Proof of the statement that the upper and lower bounds hold simultaneously (require no union bound) is also left as an exercise.

By refined Pinsker's inequality in Corollary 2.19, for $p < q$

$$\mathrm{kl}(p\|q) \geq (q-p)^2/(2q). \tag{3.24}$$

By PAC-Bayes-kl inequality, Theorem 3.26, with probability greater than $1-\delta$ for all ρ simultaneously

$$\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h,S)\right]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) \leq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}.$$

By application of inequality (3.24), the above inequality can be relaxed to

$$\mathbb{E}_\rho\left[L(h)\right] - \mathbb{E}_\rho\left[\hat{L}(h,S)\right] \leq \sqrt{2\mathbb{E}_\rho\left[L(h)\right]\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}}. \tag{3.25}$$

We have that

$$\min_{\lambda:\lambda>0}\left(\lambda x + \frac{y}{\lambda}\right) = 2\sqrt{xy}$$

(we leave this statement as a simple exercise). Thus, $\sqrt{xy} \leq \frac{1}{2}\left(\lambda x + \frac{y}{\lambda}\right)$ for all $\lambda > 0$ and by applying this inequality to (3.25) we have that with probability at least $1-\delta$ for all ρ and $\lambda > 0$

$$\mathbb{E}_\rho\left[L(h)\right] - \mathbb{E}_\rho\left[\hat{L}(h,S)\right] \leq \frac{\lambda}{2}\mathbb{E}_\rho\left[L(h)\right] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\lambda n}.$$

By changing sides

$$\left(1-\frac{\lambda}{2}\right)\mathbb{E}_\rho\left[L(h)\right] \leq \mathbb{E}_\rho\left[\hat{L}(h,S)\right] + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{\lambda n}.$$

For $\lambda < 2$ we can divide both sides by $\left(1-\frac{\lambda}{2}\right)$ and obtain the theorem statement. □

3.8.5 Alternating Minimization of PAC-Bayes-λ Bound

We use the term *PAC-Bayes-λ bound* to refer to the right hand side of PAC-Bayes-λ inequality. A great advantage of PAC-Bayes-λ bound is that it can be conveniently minimized by alternating minimization with respect to ρ and λ. Since $\mathbb{E}_\rho\left[\hat{L}(h,S)\right]$ is linear in ρ and $\mathrm{KL}(\rho\|\pi)$ is convex in ρ (Cover and Thomas, 2006), for a fixed λ the bound is convex in ρ and the minimum is achieved by

$$\rho(h) = \frac{\pi(h)e^{-\lambda n\hat{L}(h,S)}}{\mathbb{E}_\pi\left[e^{-\lambda n\hat{L}(h',S)}\right]}, \tag{3.26}$$

where $\mathbb{E}_\pi\left[e^{-\lambda n\hat{L}(h',S)}\right]$ is a convenient way of writing the normalization factor, which covers continuous and discrete hypothesis spaces in a unified notation. In the discrete case, which will be of main interest for us, $\mathbb{E}_\pi\left[e^{-\lambda n\hat{L}(h',S)}\right] = \sum_{h'\in\mathcal{H}} \pi(h')e^{-\lambda n\hat{L}(h',S)}$. We leave a proof of the statement that (3.26) defines ρ which achieves the minimum of the bound as an exercise to the reader. Furthermore, for $t \in (0,1)$ and $a, b \geq 0$ the function $\frac{a}{1-t} + \frac{b}{t(1-t)}$ is convex in t (Tolstikhin and Seldin, 2013) and, therefore, for a fixed ρ the right hand side of inequality (3.22) is convex in λ for $\lambda \in (0,2)$ and the minimum is achieved by

$$\lambda = \frac{2}{\sqrt{\frac{2n\mathbb{E}_\rho\left[\hat{L}(h,S)\right]}{\left(\mathrm{KL}(\rho\|\pi)+\ln\frac{n+1}{\delta}\right)}+1}+1}. \tag{3.27}$$

Note that the optimal value of λ is smaller than 1. Alternating application of update rules (3.26) and (3.27) monotonously decreases the bound, and thus converges.

We note that while the right hand side of inequality (3.22) is convex in ρ for a fixed λ and convex in λ for a fixed ρ, it is not simultaneously convex in ρ and λ. Joint convexity would have been a sufficient, but it is not a necessary condition for convergence of alternating minimization to the global minimum of the bound. Thiemann et al. (2017) provide sufficient conditions under which the procedure converges to the global minimum, as well as examples of situations where this does not happen.

3.8.6 Construction of a Hypothesis Space for PAC-Bayes-λ

If $\mathcal{H}$ is infinite, computation of the partition function (the denominator in (3.26)) is intractable. This could be resolved by parametrization of ρ (for example, restriction of ρ to a Gaussian posterior), but, as we have already mentioned, this may break the convexity of the bound in ρ. Fortunately, things get easy when $\mathcal{H}$ is finite. The crucial step is to construct a sufficiently powerful finite hypothesis space $\mathcal{H}$. One possibility that we consider here is to construct $\mathcal{H}$ by training m hypotheses, where each hypothesis is trained on r random points from S and validated on the remaining $n-r$ points. This construction resembles a cross-validation split of the data. However, in cross-validation r is typically large (close to n) and validation sets are non-overlapping. The approach considered here works for any r and has additional computational advantages when r is small. We do not require validation sets to be non-overlapping and overlaps between training sets are allowed. Below we describe the construction more formally.

Let $h \in \{1,\ldots,m\}$ index the hypotheses in $\mathcal{H}$. Let S_h denote the training set of h and $S \setminus S_h$ the validation set. S_h is a subset of r points from S, which are selected independently of their values (for example, subsampled randomly or picked according to a predefined partition of the data). We define the validation error of h by $\hat{L}^{\mathrm{val}}(h,S) = \frac{1}{n-r}\sum_{(X,Y)\in S\setminus S_h} \ell(h(X),Y)$. Note that the validation errors are $(n-r)$ i.i.d. random variables with bias $L(h)$ and, therefore, for $f(h,S) = (n-r)\,\mathrm{kl}(\hat{L}^{\mathrm{val}}(h,S)\|L(h))$ we have $\mathbb{E}_S\left[e^{f(h,S)}\right] \leq (n-r)+1$. The following result is a straightforward adaptation of Theorem 3.31 to this setting (we leave the proof as an exercise to the reader).

Theorem 3.32 (PAC-Bayesian Aggregation)**.** *Let S be a sample of size n. Let $\mathcal{H}$ be a set of m hypotheses, where each $h \in \mathcal{H}$ is trained on r points from S selected independently of the composition of S. For any probability distribution π over $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability greater than $1-\delta$ over a random draw of a sample S, for all distributions ρ over $\mathcal{H}$ and $\lambda \in (0,2)$ simultaneously:*

$$\mathbb{E}_\rho\left[L(h)\right] \leq \frac{\mathbb{E}_\rho\left[\hat{L}^{\mathrm{val}}(h,S)\right]}{1-\frac{\lambda}{2}} + \frac{\mathrm{KL}(\rho\|\pi)+\ln\frac{(n-r)+1}{\delta}}{\lambda\left(1-\frac{\lambda}{2}\right)(n-r)}. \tag{3.28}$$

It is natural, but not mandatory to select a uniform prior $\pi(h) = 1/m$. The bound in equation (3.28) can be minimized by alternating application of the update rules in equations (3.26) and (3.27) with n being replaced by $n-r$ and $\hat{L}$ by $\hat{L}^{\mathrm{val}}$. For evaluation of the empirical performance of this learning approach see Thiemann et al. (2017).

3.9 PAC-Bayesian Analysis of Ensemble Classifiers

So far in this chapter we have discussed various methods of selection of classifiers from a hypothesis set $\mathcal{H}$. We now turn to *aggregation* of predictions of multiple classifiers through a *weighted majority*

vote. The power of the majority vote is in the "cancellation of errors" effect: *if* predictions of different classifiers are uncorrelated and they all predict better than a random guess (meaning that $L(h) < 1/2$), the errors tend to cancel out. This can be compared to a consultation of medical experts, which tends to predict better than the best expert in the set. Most machine learning competitions are won by strategies that aggregate predictions of multiple classifiers. The assumptions that the errors are uncorrelated and the predictions are better than random are important. For example, if we have three hypotheses with $L(h) = p$ and independent errors, the probability that a uniform majority vote MV_u makes an error equals the probability that at least two out of the three hypotheses make an error. You are welcome to verify that in this case for $p \leq 1/2$ we have $L(\mathrm{MV}_u) \leq \mathbb{E}_u\left[L(h)\right]$, where u is the uniform distribution. If the errors are correlated, it can be shown that $L(\mathrm{MV}_\rho)$ can be larger than $\mathbb{E}_\rho\left[L(h)\right]$, but as we show below it is never larger than $2\mathbb{E}_\rho\left[L(h)\right]$. The reader is welcome to construct an example, where $L(\mathrm{MV}_u) > \mathbb{E}_u\left[L(h)\right]$.

3.9.1 Ensemble Classifiers and Weighted Majority Vote

We now turn to some formal definitions. Ensemble classifiers predict by taking a weighted aggregation of predictions by hypotheses from $\mathcal{H}$. In multi-class prediction (the label space $\mathcal{Y}$ is finite) ρ-weighted majority vote MV_ρ predicts

$$\mathrm{MV}_\rho(X) = \arg\max_{Y \in \mathcal{Y}} \sum_{(h \in \mathcal{H}) \wedge (h(X) = Y)} \rho(h),$$

where $\wedge$ represents the logical "and" operation and ties can be resolved arbitrarily.

In binary prediction with prediction space $h(X) \in \{\pm 1\}$ weighted majority vote can be written as

$$\mathrm{MV}_\rho(X) = \mathrm{sign}\left(\mathbb{E}_\rho\left[h(X)\right]\right),$$

where $\mathrm{sign}(x) = 1$ if $x > 0$ and -1 otherwise (the value of $\mathrm{sign}(0)$ can be defined arbitrarily). For a countable hypothesis space this becomes

$$\mathrm{MV}_\rho(X) = \mathrm{sign}\left(\sum_{h \in \mathcal{H}} \rho(h) h(X)\right).$$

3.9.2 First Order Oracle Bound for the Weighted Majority Vote

If majority vote makes an error, we know that at least a ρ-weighted half of the classifiers have made an error and, therefore, $\ell(\mathrm{MV}_\rho(X), Y) \leq \mathbb{1}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$. This observation leads to the well-known first order oracle bound for the loss of weighted majority vote.

Theorem 3.33 (First Order Oracle Bound)**.**

$$L(\mathrm{MV}_\rho) \leq 2\mathbb{E}_\rho[L(h)].$$

Proof. We have $L(\mathrm{MV}_\rho) = \mathbb{E}_D[\ell(\mathrm{MV}_\rho(X), Y)] \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5)$. By applying Markov's inequality to random variable $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ we have:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5) \leq 2\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] = 2\mathbb{E}_\rho[L(h)].$$

□

PAC-Bayesian analysis can be used to bound $\mathbb{E}_\rho[L(h)]$ in Theorem 3.33 in terms of $\mathbb{E}_\rho[\hat{L}(h, S)]$, thus turning the oracle bound into an empirical one. The disadvantage of the first order approach is that $\mathbb{E}_\rho[L(h)]$ ignores correlations of predictions, which is the main power of the majority vote.

3.9.3 Second Order Oracle Bound for the Weighted Majority Vote

Now we present a second order bound for the weighted majority vote, which is based on a second order Markov's inequality: for a non-negative random variable Z and $\varepsilon > 0$, we have $\mathbb{P}(Z \geq \varepsilon) = \mathbb{P}\big(Z^2 \geq \varepsilon^2\big) \leq \varepsilon^{-2}\mathbb{E}\left[Z^2\right]$. We define *tandem loss* of two hypotheses h and h' by

$$\ell(h(X), h'(X), Y) = \mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y).$$

The tandem loss counts an error on a sample (X, Y) only if both h and h' err on (X, Y). We define the expected tandem loss by

$$L(h, h') = \mathbb{E}_D[\mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y)].$$

The following lemma relates the expectation of the second moment of the standard loss to the expected tandem loss. We use the shorthand $\mathbb{E}_{\rho^2}[L(h, h')] = \mathbb{E}_{h\sim\rho, h'\sim\rho}[L(h, h')]$.

Lemma 3.34. *In multiclass classification*

$$\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] = \mathbb{E}_{\rho^2}[L(h, h')].$$

Proof.

$$\begin{aligned}
\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] &= \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]] \\
&= \mathbb{E}_D[\mathbb{E}_{\rho^2}[\mathbb{1}(h(X) \neq Y)\mathbb{1}(h'(X) \neq Y)]] \\
&= \mathbb{E}_D[\mathbb{E}_{\rho^2}[\mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y)]] \\
&= \mathbb{E}_{\rho^2}[\mathbb{E}_D[\mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y)]] \\
&= \mathbb{E}_{\rho^2}[L(h, h')].
\end{aligned} \tag{3.29}$$

□

A combination of second order Markov's inequality with Lemma 3.34 leads to the following result.

Theorem 3.35 (Second Order Oracle Bound)**.** *In multiclass classification*

$$L(\mathrm{MV}_\rho) \leq 4\mathbb{E}_{\rho^2}[L(h, h')]. \tag{3.30}$$

Proof. By second order Markov's inequality applied to $Z = \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]$ and Lemma 3.34:

$$L(\mathrm{MV}_\rho) \leq \mathbb{P}(\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] \geq 0.5) \leq 4\mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]^2] = 4\mathbb{E}_{\rho^2}[L(h, h')].$$

□

A Specialized Bound for Binary Classification

We provide an alternative form of Theorem 3.35, which can be used to exploit unlabeled data in binary classification. We denote the *expected disagreement* between hypotheses h and h' by $\mathbb{D}(h, h') = \mathbb{E}_D[\mathbb{1}(h(X) \neq h'(X))]$ and express the tandem loss in terms of standard loss and disagreement.

Lemma 3.36. *In binary classification*

$$\mathbb{E}_{\rho^2}[L(h, h')] = \mathbb{E}_\rho[L(h)] - \frac{1}{2}\mathbb{E}_{\rho^2}[\mathbb{D}(h, h')].$$

Proof of Lemma 3.36. Picking from (3.29), we have

$$\begin{aligned}
\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] &= \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)](1 - \mathbb{E}_\rho[(1 - \mathbb{1}(h(X) \neq Y))] \\
&= \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] - \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)]\mathbb{E}_\rho[\mathbb{1}(h(X) = Y)] \\
&= \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] - \mathbb{E}_{\rho^2}[\mathbb{1}(h(X) \neq Y \wedge h'(X) = Y)] \\
&= \mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] - \frac{1}{2}\mathbb{E}_{\rho^2}[\mathbb{1}(h(X) \neq h'(X))].
\end{aligned}$$

By taking expectation with respect to D on both sides and applying Lemma 3.34 to the left hand side, we obtain:

$$\mathbb{E}_{\rho^2}[L(h, h')] = \mathbb{E}_D[\mathbb{E}_\rho[\mathbb{1}(h(X) \neq Y)] - \frac{1}{2}\mathbb{E}_{\rho^2}[\mathbb{1}(h(X) \neq h'(X))]] = \mathbb{E}_\rho[L(h)] - \frac{1}{2}\mathbb{E}_{\rho^2}[\mathbb{D}(h, h')].$$

□

The lemma leads to the following result.

Theorem 3.37 (Second Order Oracle Bound for Binary Classification)**.** *In binary classification*

$$L(\mathrm{MV}_\rho) \leq 4\mathbb{E}_\rho[L(h)] - 2\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]. \tag{3.31}$$

Proof. The theorem follows by plugging the result of Lemma 3.36 into Theorem 3.35. □

The advantage of the alternative way of writing the bound is the possibility of using unlabeled data for estimation of $\mathbb{D}(h,h')$ in binary prediction (see also Germain et al., 2015). We note, however, that estimation of $\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')]$ has a slow convergence rate, as opposed to $\mathbb{E}_{\rho^2}[L(h,h')]$, which has a fast convergence rate. We discuss this point in Section 3.9.7.

3.9.4 Comparison of the First and Second Order Oracle Bounds

From Theorems 3.33 and 3.37 we see that in binary classification the second order bound is tighter when $\mathbb{E}_{\rho^2}[\mathbb{D}(h,h')] > \mathbb{E}_\rho[L(h)]$. Below we provide a more detailed comparison of Theorems 3.33 and 3.35 in the worst, the best, and the independent cases. The comparison only concerns the oracle bounds, whereas estimation of the oracle quantities, $\mathbb{E}_\rho[L(h)]$ and $\mathbb{E}_{\rho^2}[L(h,h')]$, is discussed in Section 3.9.7.

The worst case Since $\mathbb{E}_{\rho^2}[L(h,h')] \leq \mathbb{E}_\rho[L(h)]$ the second order bound is at most twice worse than the first order bound. The worst case happens, for example, if all hypotheses in $\mathcal{H}$ give identical predictions. Then $\mathbb{E}_{\rho^2}[L(h,h')] = \mathbb{E}_\rho[L(h)] = L(\mathrm{MV}_\rho)$ for all ρ.

The best case Imagine that $\mathcal{H}$ consists of $M \geq 3$ hypotheses, such that each hypothesis errs on $1/M$ of the sample space (according to the distribution D) and that the error regions are disjoint. Then $L(h) = 1/M$ for all h and $L(h,h') = 0$ for all $h \neq h'$ and $L(h,h) = 1/M$. For a uniform distribution ρ on $\mathcal{H}$ the first order bound is $2\mathbb{E}_\rho[L(h)] = 2/M$ and the second order bound is $4\mathbb{E}_{\rho^2}[L(h,h')] = 4/M^2$ and $L(\mathrm{MV}_\rho) = 0$. In this case the second order bound is an order of magnitude tighter than the first order.

The independent case Assume that all hypotheses in $\mathcal{H}$ make independent errors and have the same error rate, $L(h) = L(h')$ for all h and h'. Then for $h \neq h'$ we have $L(h,h') = \mathbb{E}_D[\mathbb{1}(h(X) \neq Y \wedge h'(X) \neq Y)] = \mathbb{E}_D[\mathbb{1}(h(X) \neq Y)\mathbb{1}(h'(X) \neq Y)] = \mathbb{E}_D[\mathbb{1}(h(X) \neq Y)]\mathbb{E}_D[\mathbb{1}(h'(X) \neq Y)] = L(h)^2$ and $L(h,h) = L(h)$. For a uniform distribution ρ the second order bound is $4\mathbb{E}_{\rho^2}[L(h,h')] = 4(L(h)^2 + \frac{1}{M}L(h)(1-L(h)))$ and the first order bound is $2\mathbb{E}_\rho[L(h)] = 2L(h)$. Assuming that M is large, so that we can ignore the second term in the second order bound, we obtain that it is tighter for $L(h) < 1/2$ and looser otherwise. The former is the interesting regime, especially in binary classification.

3.9.5 Second Order PAC-Bayesian Bounds for the Weighted Majority Vote

Now we provide an empirical bound for the weighted majority vote. We define the *empirical tandem loss*

$$\hat{L}(h,h',S) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}(h(X_i) \neq Y_i \wedge h'(X_i) \neq Y_i)$$

and provide a bound on the expected loss of ρ-weighted majority vote in terms of the empirical tandem losses.

Theorem 3.38. *For any probability distribution π on $\mathcal{H}$ that is independent of S and any $\delta \in (0,1)$, with probability at least $1-\delta$ over a random draw of S, for all distributions ρ on $\mathcal{H}$ and all $\lambda \in (0,2)$ simultaneously:*

$$L(\mathrm{MV}_\rho) \leq 4\left(\frac{\mathbb{E}_{\rho^2}[\hat{L}(h,h',S)]}{1-\lambda/2} + \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta)}{\lambda(1-\lambda/2)n}\right).$$

Proof. The theorem follows by using the bound in equation (3.22) to bound $\mathbb{E}_{\rho^2}[L(h,h')]$ in Theorem 3.35. We note that $\mathrm{KL}(\rho^2\|\pi^2) = 2\,\mathrm{KL}(\rho\|\pi)$ (Germain et al., 2015, Page 814). □

It is also possible to use PAC-Bayes-kl to bound $\mathbb{E}_{\rho^2}[L(h,h')]$ in Theorem 3.35, which actually gives a tighter bound, but the bound in Theorem 3.38 is more convenient for minimization. We refer the reader to Masegosa et al. (2020) for a procedure for bound minimization.

A specialized bound for binary classification

We define the *empirical disagreement*

$$\hat{\mathbb{D}}(h, h', S') = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}(h(X_i) \neq h'(X_i)),$$

where $S' = \{X_1, \ldots, X_m\}$. The set S' may overlap with the labeled set S, however, S' may include additional unlabeled data. The following theorem bounds the loss of weighted majority vote in terms of empirical disagreements. Due to possibility of using unlabeled data for estimation of disagreements in the binary case, the theorem has the potential of yielding a tighter bound when a considerable amount of unlabeled data is available.

Theorem 3.39. *In binary classification, for any probability distribution π on $\mathcal{H}$ that is independent of S and S' and any $\delta \in (0, 1)$, with probability at least $1 - \delta$ over a random draw of S and S', for all distributions ρ on $\mathcal{H}$ and all $\lambda \in (0, 2)$ and $\gamma > 0$ simultaneously:*

$$L(\mathrm{MV}_\rho) \leq 4 \left(\frac{\mathbb{E}_\rho[\hat{L}(h, S)]}{1 - \lambda/2} + \frac{\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{n}/\delta)}{\lambda(1 - \lambda/2)n} \right) - 2 \left((1 - \gamma/2)\mathbb{E}_{\rho^2}[\hat{\mathbb{D}}(h, h', S')] - \frac{2\,\mathrm{KL}(\rho\|\pi) + \ln(4\sqrt{m}/\delta)}{\gamma m} \right).$$

Proof. The theorem follows by using the upper bound in equation (3.22) to bound $\mathbb{E}_\rho[L(h)]$ and the lower bound in equation (3.23) to bound $\mathbb{E}_{\rho^2}[\mathbb{D}(h, h')]$ in Theorem 3.37. We replace δ by $\delta/2$ in the upper and lower bound and take a union bound over them. □

Using PAC-Bayes-kl to bound $\mathbb{E}_\rho[L(h)]$ and $\mathbb{E}_{\rho^2}[\mathbb{D}(h, h')]$ in Theorem 3.37 gives a tighter bound, but the bound in Theorem 3.39 is more convenient for minimisation. We refer to Masegosa et al. (2020) for a procedure for bound minimization.

3.9.6 Ensemble Construction

It is possible to use the same procedure as in Section 3.8.6 to construct an ensemble. Tandem losses can then be estimated on overlaps of validation sets, $(S \setminus S_h) \cap (S \setminus S_{h'})$. The sample size in Theorem 3.38 should then be replaced by $\min_{h,h'} |(S \setminus S_h) \cap (S \setminus S_{h'})|$.

3.9.7 Comparison of the Empirical Bounds

We provide a high-level comparison of the empirical first order bound (FO), the empirical second order bound based on the tandem loss (TND, Theorem 3.38), and the new empirical second order bound based on disagreements (DIS, Theorem 3.39). The two key quantities in the comparison are the sample size n in the denominator of the bounds and fast and slow convergence rates for the standard (first order) loss, the tandem loss, and the disagreements. Tolstikhin and Seldin (2013) have shown that if we optimize λ for a given ρ, the PAC-Bayes-λ bound in equation (3.22) can be written as

$$\mathbb{E}_\rho[L(h)] \leq \mathbb{E}_\rho[\hat{L}(h, S)] + \sqrt{\frac{2\mathbb{E}_\rho[\hat{L}(h, S)]\,(\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta))}{n}} + \frac{2\,(\mathrm{KL}(\rho\|\pi) + \ln(2\sqrt{n}/\delta))}{n}.$$

This form of the bound, also used by McAllester (2003), is convenient for explanation of fast and slow rates. If $\mathbb{E}_\rho[\hat{L}(h, S)]$ is large, then the middle term on the right hand side dominates the complexity and the bound decreases at the rate of $1/\sqrt{n}$, which is known as a *slow rate*. If $\mathbb{E}_\rho[\hat{L}(h, S)]$ is small, then the last term dominates and the bound decreases at the rate of $1/n$, which is known as a *fast rate*.

FO **vs.** TND The advantage of the FO bound is that the validation sets $S \setminus S_h$ available for estimation of the first order losses $\hat{L}(h, S_h)$ are larger than the validation sets $(S \setminus S_h) \cap (S \setminus S_{h'})$ available for estimation of the tandem losses. Therefore, the denominator $n_{\texttt{min}} = \min_h |S \setminus S_h|$ in the FO bound is larger than the denominator $n_{\texttt{min}} = \min_{h,h'} |(S \setminus S_h) \cap (S \setminus S_{h'})|$ in the TND bound. The TND disadvantage can be reduced by using data splits with large validation sets $S \setminus S_h$ and small training sets S_h, as long as small training sets do not overly impact the quality of base classifiers h. Another advantage of the FO bound is that its complexity term has $\mathrm{KL}(\rho\|\pi)$, whereas the TND bound has $2\,\mathrm{KL}(\rho\|\pi)$. The advantage of the TND bound is that $\mathbb{E}_{\rho^2}[L(h, h')] \leq E_\rho[L(h)]$ and, therefore, the convergence rate of the tandem loss is typically faster than the convergence rate of the first order loss. The interplay of the estimation advantages and disadvantages, combined with the advantages and disadvantages of the underlying oracle bounds discussed in Section 3.9.4, depends on the data and the hypothesis space.

TND **vs.** DIS The advantage of the DIS bound relative to the TND bound is that in presence of a large amount of unlabeled data the disagreements $\mathbb{D}(h, h')$ can be tightly estimated (the denominator m is large) and the estimation complexity is governed by the first order term, $\mathbb{E}_\rho[L(h)]$, which is "easy" to estimate, as discussed above. However, the DIS bound has two disadvantages. A minor one is its reliance on estimation of two quantities, $\mathbb{E}_\rho[L(h)]$ and $\mathbb{E}_{\rho^2}[\mathbb{D}(h, h')]$, which requires a union bound, e.g., replacement of δ by $\delta/2$. A more substantial one is that the disagreement term is desired to be large, and thus has a slow convergence rate. Since slow convergence rate relates to fast convergence rate as $1/\sqrt{n}$ to $1/n$, as a rule of thumb the DIS bound is expected to outperform TND only when the amount of unlabeled data is at least quadratic in the amount of labeled data, $m > n^2$.

For experimental comparison of the bounds and further details we refer the reader to Masegosa et al. (2020).

Chapter 4

Supervised Learning - Regression

In this chapter we consider the regression problem, which is another special case of supervised learning with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$.

4.1 Linear Least Squares

Linear regression with square loss $\ell(Y', Y) = (Y' - Y)^2$ is also known as linear least squares. Let $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$ be our sample. We are looking for a prediction rule of a form $h(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, where $\mathbf{w}^T\mathbf{x}$ is the dot-product (also known as the inner product) between a vector $\mathbf{w} \in \mathbb{R}^d$ and a data point $\mathbf{x} \in \mathbb{R}^d$. We will use $\mathbf{w}$ to denote the above prediction rule. Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ be a matrix holding $\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T$ as its rows

$$\mathbf{X} = \begin{pmatrix} — \ \mathbf{x}_1^T \ — \\ \vdots \\ — \ \mathbf{x}_n^T \ — \end{pmatrix}$$

and let $\mathbf{y} = (y_1, \ldots, y_n)^T$ be the vector of labels. We are looking for $\mathbf{w}$ that minimizes the empirical loss $\hat{L}(\mathbf{w}, S) = \sum_{i=1}^n \ell(\mathbf{w}^T\mathbf{x}_i, y_i) = \sum_{i=1}^n \left(\mathbf{w}^T\mathbf{x}_i - y_i\right)^2 = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$.

When the number of constrains n (the number of points in S) is larger than the number of unknowns d (the number of entries in $\mathbf{w}$), most often the linear system $\mathbf{X}\mathbf{w} = \mathbf{y}$ has no solutions (unless $\mathbf{y}$ by chance falls in the linear span of the columns of $\mathbf{X}$). Therefore, we are looking for the best approximation of $\mathbf{y}$ by a linear combination of the columns of $\mathbf{X}$, which means that we are looking for a *projection* of $\mathbf{y}$ onto the column space of $\mathbf{X}$. There are two ways to define projections, analytical and algebraic, which lead to two ways of solving the problem. In the analytical formulation the projection is a point of a form $\mathbf{X}\mathbf{w}$ that has minimal distance to $\mathbf{y}$. In the algebraic formulation the projection is a vector $\mathbf{X}\mathbf{w}$ that is perpendicular to the remainder $\mathbf{y} - \mathbf{X}\mathbf{w}$. We present both ways in detail below.

4.1.1 Analytical Approach

We are looking for

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 = \min_{\mathbf{w}} (\mathbf{X}\mathbf{w} - \mathbf{y})^T(\mathbf{X}\mathbf{w} - \mathbf{y}) = \min_{\mathbf{w}} \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^T\mathbf{y}.$$

By taking a derivative of the above and equating it to zero we have[1]

$$\frac{d(\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{y}^T\mathbf{X}\mathbf{w} + \mathbf{y}^T\mathbf{y})}{d\mathbf{w}} = 2\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{X}^T\mathbf{y} = 0.$$

Which gives

$$\mathbf{X}^T\mathbf{X}\mathbf{w} = \mathbf{X}^T\mathbf{y}.$$

If we assume that the *columns* of $\mathbf{X}$ are linearly independent ($dim(\mathbf{X}) = d$) then $\mathbf{X}^T\mathbf{X} \in \mathbb{R}^{d \times d}$ is invertible (see Appendix C) and we obtain

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.$$

[1] See Appendix D for details on calculation of derivatives [gradients] of multidimensional functions.

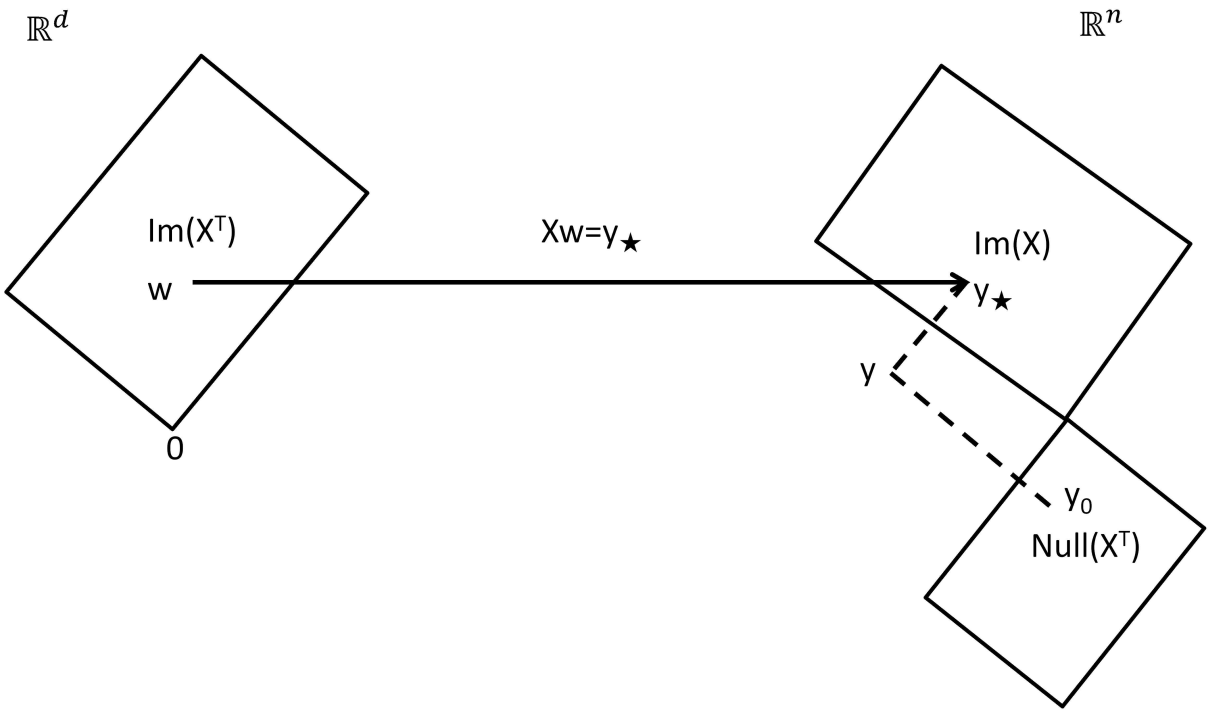


Figure 4.1: **Illustration of algebraic solution of linear least squares.**

4.1.2 Algebraic Approach - Fast Track

The projection $\mathbf{Xw}$ is a vector that is orthogonal to the remainder $\mathbf{y} - \mathbf{Xw}$ (so that $\mathbf{y}$ is a sum of the projection and the remainder, $\mathbf{y} = \mathbf{Xw} + (\mathbf{y} - \mathbf{Xw})$, and there is a right angle between the two). Two vectors are orthogonal if and only if their inner product is zero. Thus, we are looking for $\mathbf{w}$ that satisfies

$$(\mathbf{Xw})^T (\mathbf{y} - \mathbf{Xw}) = 0,$$

which is equivalent to $\mathbf{w}^T\mathbf{X}^T (\mathbf{y} - \mathbf{Xw}) = 0$. It is sufficient to find $\mathbf{w}$ that satisfies $\mathbf{X}^T (\mathbf{y} - \mathbf{Xw}) = 0$ to solve this equation, which is equivalent to $\mathbf{X}^T\mathbf{Xw} = \mathbf{X}^T\mathbf{y}$. By multiplying both sides by $\left(\mathbf{X}^T\mathbf{X}\right)^{-1}$ (which is defined, since the columns are linearly independent) we obtain a solution $\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

This solution is, actually, unique due to independence of the columns of $\mathbf{X}$. Assume there is another solution $\mathbf{w}'$, such that $\mathbf{Xw}' = \mathbf{y}$. Then $\mathbf{Xw} - \mathbf{Xw}' = \mathbf{X}(\mathbf{w} - \mathbf{w}') = 0$, but since the columns of $\mathbf{X}$ are linearly independent the only their linear combination that yields zero is the zero vector, meaning that $\mathbf{w} - \mathbf{w}' = 0$ and $\mathbf{w} = \mathbf{w}'$.

4.1.3 Algebraic Approach - Complete Picture

Linear Least Squares is a great opportunity to revisit a number of basic concepts from linear algebra. Once the complete picture is understood, the algebraic solution of the problem is just one line. We refer the reader to Appendix C for a quick review of basic concepts from linear algebra. We are looking for a solution of $\mathbf{Xw} = \mathbf{y}$, where $\mathbf{y}$ (most likely) lies outside of the column space of $\mathbf{X}$ and the equation has no solution. Therefore, the best we can do is to solve $\mathbf{Xw} = \mathbf{y}_\star$, where $\mathbf{y}_\star$ is a projection of $\mathbf{y}$ onto the column space of $\mathbf{X}$ (see Figure 4.1). We assume that $dim(\mathbf{X}) = d$ and thus the matrix $\mathbf{X}^T\mathbf{X}$ is invertible. The projection $\mathbf{y}_\star$ is then given by $\mathbf{y}_\star = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$, which means that the best we can do is to solve $\mathbf{Xw} = \mathbf{y}_\star = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$ and the solution is $\mathbf{w} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}$.

4.1.4 Using Linear Least Squares for Learning Coefficients of Non-linear Models

Linear Least Squares can be used for learning coefficients of non-linear models. For example, assume that we want to fit our data $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ (where both x_i-s and y_i-s are real numbers) with a polynomial of degree d. I.e., we want to have a model of a form $y = a_d x^d + a_{d-1} x^{d-1} + \cdots + a_1 x + a_0$. All we have to do is to map our features x_i-s into feature vectors $x_i \to (x_i^d, x_i^{d-1}, \ldots, x_i, 1)$ and apply linear least squares to the following system:

$$\begin{pmatrix} x_1^d & x_1^{d-1} & \ldots & x_1 & 1 \\ x_2^d & x_2^{d-1} & \ldots & x_2 & 1 \\ & & \vdots & & \\ x_n^d & x_n^{d-1} & \ldots & x_n & 1 \end{pmatrix} \begin{pmatrix} a_d \\ a_{d-1} \\ \vdots \\ a_1 \\ a_0 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}$$

to get the parameters vector $(a_d, a_{d-1}, \ldots, a_1, a_0)^T$.

Chapter 5

Online Learning

So far in these notes we have considered *batch* learning. In batch learning we start with some data, we analyze it, and then we "ship the result of the analysis into the world". It can be a fixed classifier h, a distribution over classifiers ρ, or anything else, the important point is that it does not change from the moment we are done with training. It takes no new information into account. This is also the reason why we had to assume that new samples come from the same distribution as the samples in the training set, because the classifier was not designed to adapt.

Online learning is a learning framework, where data collection, analysis, and application of inferred knowledge are in a perpetual loop, see Figure 5.1. Examples of problems, which fit into this framework include:

- Investment in the stock market.
- Online advertizing and personalization.
- Online routing.
- Games.
- Robotics.
- And so on ...

The recurrent nature of online learning problems makes them closely related to repeated games. They also borrow some of the terminology from the game theory, including calling the problems *games* and every "Act - Observe - Analyze" cycle a *game round*. In general, we may need online learning in the following cases:

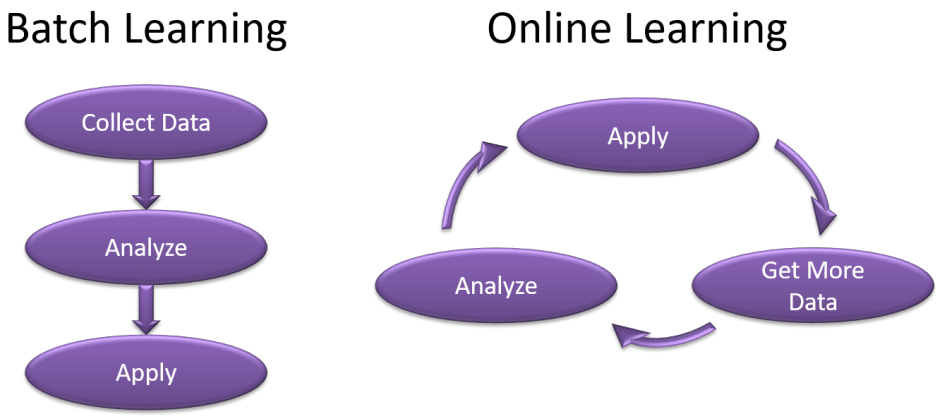


Figure 5.1: **Online learning vs. batch.**

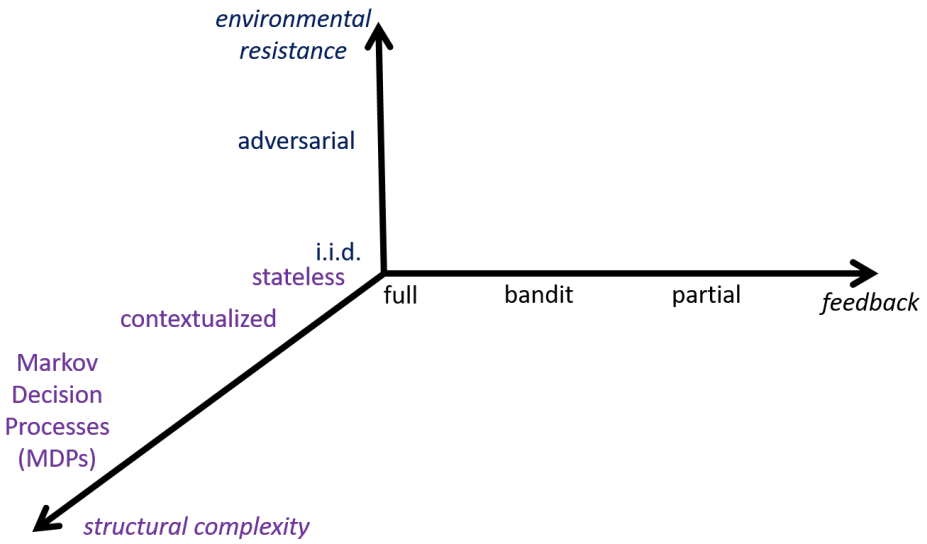


Figure 5.2: **The Space of Online Learning Problems.**

- Interactive learning: we are in a situation, where we continuously get new information and taking it into account may improve the quality of our actions. Many online applications on the Internet fall under this category.
- Adversarial or game-theoretic settings: we cannot assume that "the future behaves similarly to the past". For example, in spam filtering we cannot assume that new spam messages are generated from the same distribution as the old ones. Or, in playing chess we cannot assume that the moves of the opponent are sampled i.i.d..

As with many other problems in computer science, having loops makes things much more challenging, but also much richer and more fun.[1] For example, online learning allows to treat adversarial environments, which is impossible to do in the batch setting.

5.1 The Space of Online Learning Problems

Online learning problems are characterized by three major parameters:

1. The amount of *feedback* that the algorithm received on every round of interaction with the environment.
2. The *environmental resistance* to the algorithm.
3. The *structural complexity* of a problem.

Jointly they define *the space of online learning problems*, see Figure 5.2. It is not really a space, but a convenient way to organize the material and get initial orientation in the zoo of online learning settings. We discuss the three axes of the space with some examples below.

[1] The following quote from Robbins (1952) is interesting to read: "Until recently, statistical theory has been restricted to the design and analysis of sampling experiments in which the size and composition of the samples are completely determined before the experimentation begins. The reasons for this are partly historical, dating back to the time when the statistician was consulted, if at all, only after the experiment was over, and partly intrinsic in the mathematical difficulty of working with anything but a fixed number of independent random variables. A major advance now appears to be in the making with the creation of a theory of the *sequential design* of experiments, in which the size and composition of the samples are not fixed in advance but are functions of the observations themselves."

Feedback

Feedback refers to the amount of information that the algorithm receives on every round of interaction with the environment. The most basic forms of feedback are *full information* and *limited* (better known as *bandit*[2]) feedback.

A classical example of a full information game is investment in the stock market. On every round of this game we distribute wealth over a set of stocks and the next day we observe the rates of all stocks, which is the full information. With full information we can evaluate the quality of our investment strategy, as well as any alternative investment strategy.

A classical example of a bandit feedback game are medical treatments. We have a set of *actions* (in this case treatments), but we can only apply one treatment to a given patient. We only observe the outcome of the applied treatment, but not of any alternative treatment, thus we have limited feedback. With limited feedback we only know the quality of the selected strategy, but we cannot directly evaluate the quality of alternative strategies we could have selected. This leads to the *exploration-exploitation trade-off*, which is the trade-mark signature of online learning. The essence of the exploration-exploitation trade-off is that in order to estimate the quality of actions we have to try them out (to explore). If we explore too little, we risk missing some good actions and end up performing suboptimally. However, exploration has a cost, because trying out suboptimal actions for too long is also undesirable. The goal is to balance exploration (trying new actions) with exploitation, which is taking actions, which are currently believed to be the optimal ones. The "Act-Observe-Analyze" cycle comes into play here, because unlike in batch learning the training set is not given, but is built by the algorithm for itself: if we do not try an action we get no data from it.

There are many other problems that fall within bandit feedback framework, most notably online advertizing. A simplistic way of modeling online advertizing is assuming that there is a pool of advertizements, but on every round of the game we are only allowed to show one advertisement to a user. Since we only observe feedback for the advertisement that was presented, the problem can be formulated as an online learning problem with bandit feedback.

There are other feedback models, which we will only touch briefly. In the bandit feedback model the algorithm observes a noisy estimate of the quality of selected action, for example, whether an advertisement was clicked or not. In *partial* feedback model studied under *partial monitoring* the feedback has some relation to the action, but not necessarily its quality. For example, in dynamic pricing we only observe whether a proposed price was above or below the value of a product for a buyer, but we do not observe the maximal price we could get for the product. Bandit feedback is a special case of partial feedback, where the observation is the value. Another example is *dueling bandit* feedback, where the feedback is a relative preference over a pair of items rather than the absolute value of the items. For example, an answer to the question "Do you prefer fish or chicken?" is an example of dueling bandit feedback. Dueling bandit feedback model is used in information retrieval systems, since humans are much better in providing relative preferences rather than absolute utility values.

Environmental Resistance

Environmental resistance is concerned with how much the environment resists to the algorithm. Two classical examples are i.i.d. (a.k.a. *stochastic*) and *adversarial* environments. An example of an i.i.d. environment is the weather. It has a high degree of uncertainty, but it does not play against the algorithm. Another example of an i.i.d. environment are outcomes of medical treatments. Here also there is uncertainty in the outcomes, but the patients are not playing against the algorithm. An example of an adversarial environment is spam filtering. Here the spammers are deliberately changing distribution of the spam messages in order to outplay the spam filtering algorithm. Another classical example of an adversarial environment is the stock market. Even though the stock market does not play directly against an individual investor (assuming the investments are small), it is not stationary, because if there would be regularity in the market it would be exploited by other investors and would be gone.

The environment may also be collaborative, for example, when several agents are jointly solving a common task. Yet another example are slowly changing environments, where the parameters of a

[2] "The name derives from an imagined slot machine .... (Ordinary slot machines with one arm are one-armed bandits, since in the long run they are as effective as human bandits in separating the victim from his money.)" (Lai and Robbins, 1985)

distribution are slowly changing with time.

Structural Complexity

In structural complexity we distinguish between *stateless* problems, *contextualized* problems (or problems with state), and *Markov decision processes*. In stateless problems actions are taken without taking any additional information except the history of the outcomes into account. In contextualized problems on every round of the game the algorithm observes a context (or state) and takes an action within the observed context. An example of context is a medical record of a patient or, in the advertizing example, it could be parameters of the advertisement and the user.

Markov decision processes are concerned with processes with evolving state. The difference between contextualized problems and Markov decision processes is that in the former the actions of the algorithm do not influence the next state, whereas in the latter they do. For example, subsequent treatments of the same patient are changing his or her state and, therefore, depend on each other. In contrast, in subsequent treatments of different patients treatment of one patient does not influence the state of the next patient and, thus, can be modeled as a contextualized problem.

Markov decision processes are studied within the field of *reinforcement learning*. There is no clear cut distinction between online learning and reinforcement learning and one could be seen as a subfield of another or the other way around. But as a rule of thumb, problems involving evolution of states, such as Markov decision processes, are part of reinforcement learning and problems that do not involve evolution of states are part of online learning.

One of the challenges in Markov decision processes is *delayed feedback*. It refers to the fact that, unlike in stateless and contextualized problems, the quality of an action cannot be evaluated instantaneously. The reason is that actions are changing the state, which may lead to long-term consequences. Consider a situation of sitting in a bar, where every now and then a waiter comes and asks whether you want another beer. If you take a beer you probably feel better than if you do not, but then eventually if you take too much you will feel very bad the next morning, whereas if you do not you may feel excellent. As before, things get more challenging, but also more exciting, when there are loops in the state space.

In Markov decision processes we distinguish between *estimation* and *planning*. Estimation is the same problem as in other online learning problems - the outcomes of actions are unknown and we have to estimate them. However, in Markov decision processes even if the immediate outcomes of various actions are known, the identity of the best action in each state may still be not evident due to the long-term consequences. This problem is addressed by planning.

There are many other online learning problems, which do not fit directly into Figure 5.2, but can still be discussed in terms of feedback, environmental resistance, and structural complexity. For example, in *combinatorial bandits* the goal is to select a set of actions, potentially with some constraints, and the quality of the set is evaluated jointly. An instance of a combinatorial bandit problem is selection of a path in a graph, such as communication or transport network. In this case an action can be decomposed into sub-actions corresponding to selection of edges in the graph. The goal is to minimize the length of a path, which may correspond to the delay between the source and the target nodes. Various forms of feedback can be considered, including bandit feedback, where the total length of the path is observed; semi-bandit feedback, where the length of each of the selected edges is observed; cascading bandit feedback, where the lengths of the edges are observed in a sequence until a terminating note (e.g., a server that is down) or the target is reached; or a full information feedback, where the length of all edges is observed.

In the following sections we consider in detail a number of the most basic online learning problems.

5.2 A General Basic Setup

We start with four most basic problems in online learning, *prediction with expert advice*, *stochastic multiarmed bandits*, and *adversarial multiarmed bandits*. Prediction with expert advice refers to the adversarial version of the problem, but in the home assignment you will analyze its stochastic counterpart, which gives the fourth problem. All four are stateless problems and correspond to the four red crosses in Figure 5.3. We provide a general setup that encompasses all four problems and then specialize it. We

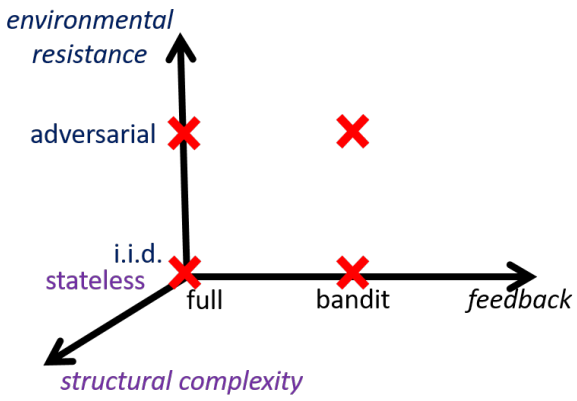


Figure 5.3: **The four basic online learning problems.**

are given a $K \times \infty$ matrix of losses ℓ_t^a, where $t \in \{1, 2, \dots\}$ and $a \in \{1, \dots, K\}$ and $\ell_t^a \in [0, 1]$.

$$\text{Losses} \begin{matrix} \ell_1^1, & \ell_2^1, & \dots & \ell_t^1, & \dots \\ \vdots & \vdots & \dots & \vdots & \dots \\ \ell_1^a, & \ell_2^a, & \dots & \ell_t^a, & \dots \\ \vdots & \vdots & \dots & \vdots & \dots \\ \ell_1^K, & \ell_2^K, & \dots & \ell_t^K, & \dots \end{matrix}$$

$$\longrightarrow \text{time}$$

The matrix is fixed before the game starts, but not revealed to the algorithm. There are two ways to generate the matrix, which are specified after the definition of the game protocol.

Game Protocol

For $t = 1, 2, \dots$:

1. Pick a row A_t
2. Suffer $\ell_t^{A_t}$
3. Observe ... [the observations are defined below]

Definition of the four games There are two common ways to generate the matrix of losses. The first is to sample ℓ_t^a-s independently, so that the mean of the losses in each row is fixed, $\mathbb{E}[\ell_t^a] = \mu(a)$. The second is to generate ℓ_t^a-s arbitrarily. The second model of generation of losses is known as an *oblivious adversary*, since the generation happens before the game starts and thus does not take actions of the algorithm into account.[3]

There are also two common ways to define the observations. After picking a row in round t the algorithm may observe either the full column $\ell_t^1, \dots, \ell_t^K$ or just the selected entry $\ell_t^{A_t}$. Jointly the two ways of generating the matrix of losses and the two ways of defining the observations generate four variants of the game.

Matrix generation \ Observations	Observe $\ell_t^1, \dots, \ell_t^K$	Observe $\ell_t^{A_t}$
ℓ_t^a-s are sampled i.i.d. with $\mathbb{E}[\ell_t^a] = \mu(a)$	I.I.D. Prediction with expert advice	Stochastic multiarmed bandits
ℓ_t^a are selected arbitrarily (by an adversary)	Prediction with expert advice (adversarial)	Adversarial multiarmed bandits

[3]It is also possible to consider an *adaptive adversary*, which generates losses as the game proceeds and takes past actions of the algorithm into account. We do not discuss this model in the lecture notes.

Performance Measure The goal of the algorithm is to play so that the loss it suffers will not be significantly larger than the loss of the best row in hindsight. There are several ways to formalize this goal. The basic performance measure is the *regret* defined by

$$R_T = \sum_{t=1}^{T} \ell_t^{A_t} - \min_a \sum_{t=1}^{T} \ell_t^a.$$

In adversarial problems we analyze the *expected regret*[4] defined by

$$\mathbb{E}\left[R_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \mathbb{E}\left[\min_a \sum_{t=1}^{T} \ell_t^a\right].$$

If the sequence of losses is deterministic we can remove the second expectation and obtain a slightly simpler expression

$$\mathbb{E}\left[R_T\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_a \sum_{t=1}^{T} \ell_t^a.$$

In stochastic problems we analyze the *pseudo regret* defined by

$$\bar{R}_T = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - \min_a \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^a\right] = \mathbb{E}\left[\sum_{t=1}^{T} \ell_t^{A_t}\right] - T \min_a \mu(a).$$

Note that since for random variables X and Y we have $\mathbb{E}\left[\min\left\{X, Y\right\}\right] \leq \min\left\{\mathbb{E}\left[X\right], \mathbb{E}\left[Y\right]\right\}$ [it is recommended to verify this identity], we have $\bar{R}_T \leq \mathbb{E}\left[R_T\right]$. A reason to consider pseudo regret in the stochastic setting is that we can get bounds of order $\ln T$ on the pseudo regret (so called "logarithmic" regret bounds), whereas the fluctuations of $\sum_{t=1}^{T} \ell_t^a$ are of order $\sqrt{T}$ (when we sample T random variables, the deviation of $\sum_{t=1}^{T} \ell_t^a$ from the expectation $T\mu(a)$ is of order $\sqrt{T}$). Thus, it is impossible to get logarithmic bounds for the expected regret.

Explanation of the Names In the complete definition of prediction with expert advice game on every round of the game the player gets an advice from K experts and then takes an action, which may be a function of the advice and the player, as well as the experts, suffer a loss depending on the action taken. Hence the name, prediction with expert advice. If we restrict the actions of the player to following the advice of a single expert, then from the perspective of the playing strategy the actual advice does not matter and it is only the loss that defines the strategy. We consider the restricted setting, because it allows to highlight the relation with multiarmed bandits.

The name multiarmed bandits comes from the analogy with slot machines, which are one-armed bandits. In this game actions are the "arms" of a slot machine.

Losses vs. Rewards In some games it is more natural to consider rewards (also called gains) rather than losses. In fact, in the literature on stochastic problems it is more popular to work with rewards, whereas in the literature on adversarial problems it is more popular to work with losses. There is a simple transformation $r = 1 - \ell$, which brings a losses game into a gains game and the other way around. Interestingly, in the adversarial setting working with losses leads to tighter and simpler results. In the stochastic setting the choice does not matter.

5.3 I.I.D. (stochastic) Multiarmed Bandits

In this section we consider multiarmed bandit game, where the outcomes are generated i.i.d. with fixed, but unknown means. In this game there is no difference between working with losses or rewards, and since most of the literature is based on games with rewards we are going to use rewards in order to be consistent. The treatment of losses is identical - see Seldin (2015).

[4]It is also possible to analyze the regret, but we do not do it here.

Notations We are given a $K \times \infty$ matrix of rewards (or gains) r_t^a, where $t \in \{1, 2, \dots\}$ and $a \in \{1, \dots, K\}$.

$$\begin{array}{c} \textit{Action rewards} \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{cccc} r_1^1, & r_2^1, & \cdots & r_t^1, & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \\ r_1^a, & r_2^a, & \cdots & r_t^a, & \cdots \\ \vdots & \vdots & \cdots & \vdots & \cdots \\ r_1^K, & r_2^K, & \cdots & r_t^K, & \cdots \\ & \longrightarrow & & & \\ & \textit{time} & & & \end{array}$$

We assume that r_t^a-s are in $[0, 1]$ and that they are generated independently, so that $\mathbb{E}[r_t^a] = \mu(a)$. We use $\mu^* = \max_a \mu(a)$ to denote the expected reward of an optimal action and $\Delta(a) = \mu^* - \mu(a)$ to denote the *suboptimality gap* (or simply the *gap*) of action a. We use $a^* = \arg\max_a \mu(a)$ to denote *a best action* (note that there may be more than one best action, in such case let a^* be any of them).

Game Definition

For $t = 1, 2, \dots$:

1. Pick a row A_t
2. Observe & accumulate $r_t^{A_t}$

Performance Measure The performance is measured by *pseudo regret* defined by

$$\begin{aligned} \bar{R}_T &= \max_a \mathbb{E}\left[\sum_{t=1}^T r_t^a\right] - \mathbb{E}\left[\sum_{t=1}^T r_t^{A_t}\right] \\ &= T\mu^* - \mathbb{E}\left[\sum_{t=1}^T r_t^{A_t}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \left(\mu^* - r_t^{A_t}\right)\right] \\ &= \mathbb{E}\left[\sum_{t=1}^T \Delta(A_t)\right] \\ &= \mathbb{E}\left[\sum_a N_T(a)\Delta(a)\right] \\ &= \sum_a \Delta(a)\mathbb{E}\left[N_T(a)\right], \end{aligned}$$

where we use $N_t(a)$ to denote the number of times action a was played up to round t. Note that in the i.i.d. setting the performance of an algorithm is compared to the best action in expectation ($\max_a \mathbb{E}\left[\sum_{t=1}^T r_t^a\right]$), whereas in the adversarial setting the performance of an algorithm is compared to the best action in hindsight ($\min_a \sum_{t=1}^T \ell_t^a$).

Exploration-exploitation trade-off: A simple approach I.i.d. multiarmed bandits is the simplest problem where we face the exploration-exploitation trade-off. In general, the goal is to play a best arm on all the rounds, but since the identity if the best arm is unknown it has to be identified first. In order to identify a best arm we need to explore all the arms. However, rounds used for exploration of suboptimal

arms increase the regret (through the $N_t(a)\Delta(a)$ term). At the same time, too greedy exploration may lead to confusion between a best and a suboptimal arm, which may eventually lead to even higher regret when we start exploiting a wrong arm. So let us make a first attempt to quantify this trade-off. Assume that we know time horizon T and we start with εT exploration rounds followed by $(1-\varepsilon)T$ exploitation rounds (where we play what we believe to be a best arm). Also assume that we have just two actions and we know that for $a \neq a^*$ we have $\Delta(a) = \Delta$. The only thing we do not know is which of the two actions is the best. So how should we set ε?

Let $\delta(\varepsilon)$ denote the probability that we misidentify the best arm at the end of the exploration period. The pseudo regret can be bounded by:

$$\bar{R}_T \leq \frac{1}{2}\Delta\varepsilon T + \delta(\varepsilon)\Delta(1-\varepsilon)T \leq \frac{1}{2}\Delta\varepsilon T + \delta(\varepsilon)\Delta T = \left(\frac{1}{2}\varepsilon + \delta(\varepsilon)\right)\Delta T,$$

where the first term is a bound on the pseudo regret during the exploration phase and the second term is a bound on the pseudo regret during the exploitation phase in case we select a wrong arm at the end of the exploration phase. Now what is $\delta(\varepsilon)$? Let $\hat{\mu}_t(a)$ denote the empirical mean of observed rewards of arm a up to round t. For the exploitation phase it is natural to select the arm that maximizes $\hat{\mu}_{\varepsilon T}(a)$ at the end of the exploration phase. Therefore:

$$\begin{aligned}
\delta(\varepsilon) &= \mathbb{P}(\hat{\mu}_{\varepsilon T}(a) \geq \hat{\mu}_{\varepsilon T}(a^*)) \\
&\leq \mathbb{P}\left(\hat{\mu}_{\varepsilon T}(a) \geq \mu(a) + \frac{1}{2}\Delta\right) + \mathbb{P}\left(\hat{\mu}_{\varepsilon T}(a^*) \leq \mu^* - \frac{1}{2}\Delta\right) \\
&\leq 2e^{-2\frac{\varepsilon T}{2}\left(\frac{1}{2}\Delta\right)^2} = 2e^{-\varepsilon T\Delta^2/4},
\end{aligned}$$

where the last line is by Hoeffding's inequality. By substituting this back into the regret bound we obtain:

$$\bar{R}_T \leq \left(\frac{1}{2}\varepsilon + 2e^{-\varepsilon T\Delta^2/4}\right)\Delta T.$$

In order to minimize $\frac{1}{2}\varepsilon + 2e^{-\varepsilon T\Delta^2/4}$ we take a derivative and equate it to zero, which leads to $\varepsilon = \frac{\ln(T\Delta^2)}{T\Delta^2/4}$. It is easy to check that the second derivative is positive, confirming that this is the minimum. Note that ε must be non-negative, so strictly speaking we have $\varepsilon = \max\left\{0, \frac{\ln(T\Delta^2)}{T\Delta^2/4}\right\}$. If we substitute this back into the regret bound we obtain:

$$\bar{R}_T \leq \max\left\{\Delta T, \left(\frac{2\ln(T\Delta^2)}{T\Delta^2} + 2e^{-\ln(T\Delta^2)}\right)\Delta T\right\} = \max\left\{\Delta T, \frac{2\ln(T\Delta^2)}{\Delta} + \frac{2}{\Delta}\right\}.$$

Note that the number of exploration rounds is $\varepsilon T = \max\left\{0, \frac{\ln(T\Delta^2)}{\Delta^2/4}\right\}$.

Put attention that the regret bound is larger when Δ is small. Although intuitively when Δ is small we do not care that much about playing a suboptimal action as opposed to the case when Δ is large, problems with small Δ are actually harder and lead to larger regret. The reason is that the number of rounds that it takes to identify the best action grows with $1/\Delta^2$. Even though in each exploration round we only suffer the regret of Δ the fact that the number of exploration rounds grows with $1/\Delta^2$ makes problems with small Δ harder.

The above approach has three problems: (1) it assumes knowledge of the time horizon T, (2) it assumes knowledge of the gap Δ, and (3) if we would try to generalize it to more than one arm the length of the exploration period would depend on the smallest gap, even if there are many arms with larger gap that are much easier to eliminate. The following approach resolves all three problems.

Upper Confidence Bound (UCB) algorithm We now concider the UCB1 algorithm of Auer et al. (2002a).

The expression $U_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3\ln t}{2N_{t-1}(a)}}$ is called an *upper confidence bound*. Why? Because $U_t(a)$ upper bounds $\mu(a)$ with high probability. UCB approach follows the *optimism in the face of uncertainty principle*. That is, we take an optimistic estimate of the reward of every arm by taking the upper limit of the confidence bound. UCB1 algorithm has the following regret guarantee.

Algorithm 3 UCB1 (Auer et al., 2002a)

Initialization: Play each action once.
for $t = K+1, K+2, \ldots$ **do**

Play $A_t = \arg\max_a \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$.

end for

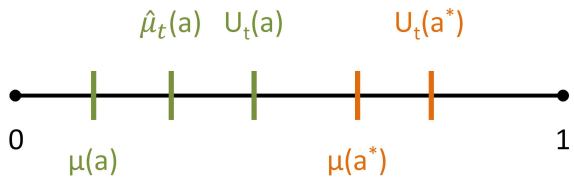


Figure 5.4: Illustration for UCB analysis.

Theorem 5.1. *For any time T the regret of UCB1 satisfies:*

$$\bar{R}_T \leq 6 \sum_{a:\Delta(a)>0} \frac{\ln T}{\Delta(a)} + \left(1 + \frac{\pi^2}{3}\right) \sum_a \Delta(a).$$

Proof. For the analysis it is convenient to have the following picture in mind - see Figure 5.4. A suboptimal arm is played when $U_t(a) \geq U_t(a^*)$. Our goal is to show that this does not happen very often. The analysis is based on the following three points, which bound the corresponding distances in Figure 5.4.

1. We show that $U_t(a^*) > \mu(a^*)$ for almost all rounds. A bit more precisely, let $F(a^*)$ be the number of rounds when $U_t(a^*) \leq \mu(a^*)$, then $\mathbb{E}[F(a^*)] \leq \frac{\pi^2}{6}$.
2. In a similar way, we show that $\hat{\mu}_t(a) < \mu(a) + \sqrt{\frac{3 \ln t}{2N_t(a)}}$ for almost all rounds. A bit more precisely, let $F(a)$ be the number of rounds when $\hat{\mu}_t(a) \geq \mu(a) + \sqrt{\frac{3 \ln t}{2N_t(a)}}$, then $\mathbb{E}[F(a)] \leq \frac{\pi^2}{6}$. (Note that this is a lower confidence bound for $\mu(a)$, or, in other words, the other side of inequality compared to Point 1.)
3. When Point 2 holds we have that $U_t(a) = \hat{\mu}_{t-1}(a) + \sqrt{\frac{3 \ln t}{2N_{t-1}(a)}} \leq \mu(a) + 2\sqrt{\frac{3 \ln t}{2N_{t-1}(a)}} = \mu(a^*) - \Delta(a) + 2\sqrt{\frac{3 \ln t}{2N_{t-1}(a)}}$.

Let us fix time horizon T and analyze what happens by time T (note that the algorithm does not depend on T). We have that for most rounds $t \leq T$:

$$U_t(a) < \mu(a^*) - \Delta(a) + \sqrt{\frac{6 \ln t}{N_{t-1}(a)}} \leq \mu(a^*) - \Delta(a) + \sqrt{\frac{6 \ln T}{N_{t-1}(a)}}$$
$$U_t(a^*) > \mu(a^*).$$

Thus, we can play a suboptimal action a only in the following cases:

- Either $\sqrt{\frac{6 \ln T}{N_{t-1}(a)}} \geq \Delta(a)$, which means that $N_{t-1}(a) \leq \frac{6 \ln T}{\Delta(a)^2}$.
- Or one of the confidence intervals in Points 1 or 2 has failed.

In other words, after a suboptimal action a has been played for $\left\lceil \frac{6 \ln T}{\Delta(a)^2} \right\rceil$ rounds it can only be played again if one of the confidence intervals fails. Therefore,

$$\mathbb{E}[N_T(a)] \leq \left\lceil \frac{6 \ln T}{\Delta(a)^2} \right\rceil + \mathbb{E}[F(a^*)] + \mathbb{E}[F(a)] \leq \frac{6 \ln T}{\Delta(a)^2} + 1 + \frac{\pi^2}{3}$$

and since $\bar{R}_T(a) = \sum_a \Delta(a)\mathbb{E}[N_T(a)]$ the result follows.

To complete the proof it is left to prove Points 1 and 2. We prove Point 1, the proof of Point 2 is identical. We start by looking at $\mathbb{P}(U_t(a^*) \le \mu(a^*)) = \mathbb{P}\Big(\hat{\mu}_{t-1}(a^*) + \sqrt{\frac{3\ln t}{2N_{t-1}(a^*)}} \le \mu(a^*)\Big) = \mathbb{P}\Big(\mu(a^*) - \hat{\mu}_{t-1}(a^*) \ge \sqrt{\frac{3\ln t}{2N_{t-1}(a^*)}}\Big)$. The delicate point is that $N_{t-1}(a^*)$ is a random variable that is not independent of $\hat{\mu}_{t-1}(a^*)$ and thus we cannot apply Hoeffding's inequality directly. Instead, we look at a series of random variables $X_1, X_2, \dots$, such that X_i-s have the same distribution as $r_t^{a^*}$-s. Let $\bar{\mu}_s = \frac{1}{s}\sum_{i=1}^{s} X_i$ be the average of the first s elements of the sequence. Then we have:

$$\begin{aligned}
\mathbb{P}\left(\mu(a^*) - \hat{\mu}_{t-1}(a^*) \ge \sqrt{\frac{3\ln t}{2N_{t-1}(a^*)}}\right) &\le \mathbb{P}\left(\exists s : \mu(a^*) - \bar{\mu}_s \ge \sqrt{\frac{3\ln t}{2s}}\right) \\
&\le \sum_{s=1}^{t} \mathbb{P}\left(\mu(a^*) - \bar{\mu}_s \ge \sqrt{\frac{3\ln t}{2s}}\right) \\
&\le \sum_{s=1}^{t} \frac{1}{t^3} = \frac{1}{t^2},
\end{aligned}$$

where in the first line we decouple $\hat{\mu}_t(a^*)$-s from $N_t(a^*)$-s via the use of $\bar{\mu}_s$-s and in the last line we apply Hoeffding's inequality (note that $3\ln t = \ln t^3$ corresponds to $\ln\frac{1}{\delta}$ in Hoeffding's inequality and thus $\delta = \frac{1}{t^3}$). Finally, we have:

$$\mathbb{E}[F(a^*)] = \sum_{t=1}^{\infty} \mathbb{P}\left(\mu(a^*) - \hat{\mu}_{t-1}(a^*) \ge \sqrt{\frac{3\ln t}{2N_{t-1}(a^*)}}\right) \le \sum_{t=1}^{\infty} \frac{1}{t^2} = \frac{\pi^2}{6}.$$

□

5.4 Prediction with Expert Advice

Notations We are given a $K \times \infty$ matrix of expert losses ℓ_t^a, where $t \in \{1, 2, \dots\}$ and $a \in \{1, \dots, K\}$.

$$\begin{array}{cc}
 & \begin{array}{cccccc} \ell_1^1, & \ell_2^1, & \cdots & \ell_t^1, & \cdots \end{array} \\
\textit{Expert Losses} & \begin{array}{cccccc} \vdots & \vdots & \cdots & \vdots & \cdots \end{array} \\
 & \begin{array}{cccccc} \ell_1^a, & \ell_2^a, & \cdots & \ell_t^a, & \cdots \end{array} \\
 & \begin{array}{cccccc} \vdots & \vdots & \cdots & \vdots & \cdots \end{array} \\
 & \begin{array}{cccccc} \ell_1^K, & \ell_2^K, & \cdots & \ell_t^K, & \cdots \end{array} \\
 & \longrightarrow \\
 & \textit{time}
\end{array}$$

Game Definition

For $t = 1, 2, \dots$:

1. Pick a row A_t
2. Observe the column $\ell_t^1, \dots, \ell_t^K$ & suffer $\ell_t^{A_t}$

Performance Measure The performance is measured by *regret*

$$R_T = \sum_{t=1}^{T} \ell_t^{A_t} - \min_a \left(\sum_{t=1}^{T} \ell_t^a\right).$$

In the notes we analyze the *expected regret* $\mathbb{E}[R_T]$.

Algorithm 4 Hedge (a.k.a. Exponential Weights), (Vovk, 1990, Littlestone and Warmuth, 1994)

Input: Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$
$\forall a : L_0(a) = 0$
for $t = 1, 2, \ldots$ **do**
  $\forall a : \; p_t(a) = \frac{e^{-\eta_t L_{t-1}(a)}}{\sum_{a'} e^{-\eta_t L_{t-1}(a')}}$
  Sample A_t according to p_t and play it
  Observe $\ell_t^1, \ldots, \ell_t^K$ and suffer $\ell_t^{A_t}$
  $\forall a : \; L_t(a) = L_{t-1}(a) + \ell_t^a$
end for

Algorithm We consider the Hedge algorithm (a.k.a. exponential weights and weighted majority) for playing this game.

Analysis We analyze the Hedge algorithm in a slightly simplified setting, where the time horizon T is known. Unknown time horizon can be handled by using the doubling trick (see home assignment) or, more elegantly, by a more careful analysis (see, e.g., Bubeck and Cesa-Bianchi (2012)).

The analysis is based on the following lemma.

Lemma 5.2. *Let $\{X_1^a, X_2^a, \ldots\}_{a \in \{1,\ldots,K\}}$ be K sequences of non-negative numbers ($X_t^a \geq 0$ for all a and t). Let $L_t(a) = \sum_{s=1}^t X_s^a$, let $L_0(a)$ be zero for all a and let $\eta > 0$. Finally, let $p_t(a) = \frac{e^{-\eta L_{t-1}(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}}$. Then:*

$$\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) X_t^a - \min_a L_T(a) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \left(X_t^a\right)^2 .$$

Proof. We define $W_t = \sum_a e^{-\eta L_t(a)}$ and study how this quantity evolves. We start with an upper bound.

$$\begin{aligned}
\frac{W_t}{W_{t-1}} &= \frac{\sum_a e^{-\eta L_t(a)}}{\sum_a e^{-\eta L_{t-1}(a)}} \\
&= \frac{\sum_a e^{-\eta X_t^a} e^{-\eta L_{t-1}(a)}}{\sum_a e^{-\eta L_{t-1}(a)}} && (5.1) \\
&= \sum_a e^{-\eta X_t^a} \frac{e^{-\eta L_{t-1}(a)}}{\sum_{a'} e^{-\eta L_{t-1}(a')}} \\
&= \sum_a e^{-\eta X_t^a} p_t(a) && (5.2) \\
&\leq \sum_a \left(1 - \eta X_t^a + \frac{1}{2}\eta^2 \left(X_t^a\right)^2\right) p_t(a) && (5.3) \\
&= 1 - \eta \sum_a X_t^a p_t(a) + \frac{\eta^2}{2} \sum_a \left(X_t^a\right)^2 p_t(a) \\
&\leq e^{-\eta \sum_a X_t^a p_t(a) + \frac{\eta^2}{2} \sum_a (X_t^a)^2 p_t(a)}, && (5.4)
\end{aligned}$$

where in (5.1) we used the fact that $L_t(a) = X_t^a + L_{t-1}(a)$, in (5.2) we used the definition of $p_t(a)$, in (5.3) we used the inequality $e^x \leq 1 + x + \frac{1}{2}x^2$, which holds for $x \leq 0$ (this is a delicate point, because the inequality does not hold for $x > 0$ and, therefore, we must check that the condition $x \leq 0$ is satisfied; it is satisfied under the assumptions of the lemma), and inequality (5.4) is based on inequality $1 + x \leq e^x$, which holds for all x.

Now we consider the ratio $\frac{W_T}{W_0}$. On the one hand:

$$\frac{W_T}{W_0} = \frac{W_1}{W_0} \times \frac{W_2}{W_1} \times \cdots \times \frac{W_T}{W_{T-1}} \leq e^{-\eta \sum_{t=1}^T \sum_a X_t^a p_t(a) + \frac{\eta^2}{2} \sum_{t=1}^T \sum_a (X_t^a)^2 p_t(a)}.$$

On the other hand:

$$\frac{W_T}{W_0} = \frac{\sum_a e^{-\eta L_T(a)}}{K} \geq \frac{\max_a e^{-\eta L_T(a)}}{K} = \frac{e^{-\eta \min_a L_T(a)}}{K},$$

where we lower-bounded the sum by its maximal element. By taking the two inequalities together and applying logarithm we obtain:

$$-\eta \min_a L_T(a) - \ln K \leq -\eta \sum_{t=1}^{T} \sum_a X_t^a p_t(a) + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_a \left(X_t^a\right)^2 p_t(a).$$

Finally, by changing sides and dividing by η we get:

$$\sum_{t=1}^{T} \sum_a X_t^a p_t(a) - \min_a L_T(a) \leq \frac{\ln K}{\eta} + \frac{\eta}{2} \sum_{t=1}^{T} \sum_a \left(X_t^a\right)^2 p_t(a)$$

□

Now we are ready to present an analysis of the Hedge algorithm.

Theorem 5.3. *The expected regret of the Hedge algorithm with a fixed learning rate η satisfies:*

$$\mathbb{E}\left[R_T\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2} T.$$

The expected regret is minimized by $\eta = \sqrt{\frac{2 \ln K}{T}}$, which leads to

$$\mathbb{E}\left[R_T\right] \leq \sqrt{2T \ln K}.$$

Proof. We note that ℓ_t^a-s are positive and apply Lemma 5.2 to obtain:

$$\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \ell_t^a - \min_a L_T(a) \leq \frac{\ln K}{\eta} + \frac{\eta^2}{2} \sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \left(\ell_t^a\right)^2.$$

Note that $\sum_a p_t(a) \ell_t^a$ is the expected loss of Hedge on round t and $\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \ell_t^a$ is the expected cumulative loss of Hedge after T rounds. Thus, the left hand side of the inequality is the expected regret of Hedge. Also note that $\ell_t^a \leq 1$ and thus $\left(\ell_t^a\right)^2 \leq 1$ and $\sum_a p_t(a) \left(\ell_t^a\right)^2 \leq 1$. Thus, $\sum_{t=1}^{T} \sum_{a=1}^{K} p_t(a) \left(\ell_t^a\right)^2 \leq T$. Altogether, we get that:

$$\mathbb{E}\left[R_T\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2} T.$$

By taking the derivative of the right hand side and equating it to zero we obtain that $-\frac{\ln K}{\eta^2} + \frac{T}{2} = 0$ and thus $\eta = \sqrt{\frac{2 \ln K}{T}}$ is an extremal point. The second derivative is $\frac{2 \ln K}{\eta^3}$ and since $\eta > 0$ it is positive. Thus, the extremal point is the minimum. □

5.4.1 Lower Bound

A lower bound for the expected regret in prediction with expert advice is based on the following construction. We draw a $K \times \infty$ matrix of losses with each loss drawn according to Bernoulli distribution with bias $1/2$. In this game the expected loss of any algorithm after T rounds is $T/2$, irrespective of what the algorithm is doing. However, the loss of the best action in hindsight is lower, because we are selecting the "best" out of K rows. For each individual row the expected loss is $T/2$, but the expectation of the minimum of the losses is lower. The reduction is quantified in the following theorem, see Cesa-Bianchi and Lugosi (2006) for a proof.

Theorem 5.4. *Let ℓ_t^a be i.i.d. Bernoulli random variables with bias $1/2$, then*

$$\lim_{T\to\infty}\lim_{K\to\infty}\frac{T/2-\mathbb{E}\left[\min_a\sum_{t=1}^T\ell_t^a\right]}{\sqrt{\frac{1}{2}T\ln K}}=1.$$

Note that the numerator in the above expression, $T/2-\mathbb{E}\left[\min_a\sum_{t=1}^T\ell_t^a\right]$, is the expectation with respect to generation of the matrix of losses of the expected regret. Thus, if the adversary generates the matrix of losses according to the construction described above, then in expectation with respect to generation of the matrix and in the limit of K and T going to infinity the expected regret cannot be smaller than $\sqrt{\frac{1}{2}T\ln K}$.

5.5 Adversarial Multiarmed Bandits

Game definition We are working with the same matrix of losses as in prediction with expert advice, but now at each round of the game we are allowed to observe only the loss of the row that we have played:
For $t=1,2,\ldots$:

1. Pick a row A_t
2. Observe & suffer $\ell_t^{A_t}$. (ℓ_t^a-s for $a\neq A_t$ remain unobserved)

Algorithm The algorithm is based on using importance-weighted estimates of the losses in the Hedge algorithm.[5]

Algorithm 5 EXP3 (Auer et al., 2002b)

Input: Learning rates $\eta_1\geq\eta_2\geq\cdots>0$
$\forall a:\tilde{L}_0(a)=0$
for $t=1,2,\ldots$ **do**
  $\forall a:\ p_t(a)=\frac{e^{-\eta_t\tilde{L}_{t-1}(a)}}{\sum_{a'}e^{-\eta_t\tilde{L}_{t-1}(a')}}$
  Sample A_t according to p_t and play it
  Observe and suffer $\ell_t^{A_t}$
  Set $\tilde{\ell}_t^a=\frac{\ell_t^a\mathbb{1}(A_t=a)}{p_t(a)}=\begin{cases}\frac{\ell_t^a}{p_t(a)}, & \text{If } A_t=a\\ 0, & \text{otherwise}\end{cases}$
  $\forall a:\ \tilde{L}_t(a)=\tilde{L}_{t-1}(a)+\tilde{\ell}_t^a$
end for

Properties of importance-weighted samples Before we analyze the EXP3 algorithm we discuss a number of important properties of importance-weighted sampling.

1. The samples $\tilde{\ell}_t^a$ are not independent in two ways. First, for a fixed t, the set $\left\{\tilde{\ell}_t^1,\ldots,\tilde{\ell}_t^K\right\}$ is dependent (if we know that one of $\tilde{\ell}_t^a$-s is non-zero, we automatically know that all the rest are zero). And second, $\tilde{\ell}_t^a$ depends on all $\tilde{\ell}_s^{a'}$ for $s<t$ and all a' since $p_t(a)$ depends on $\left\{\tilde{\ell}_s^a\right\}_{1\leq s<t,a\in\{1,\ldots,K\}}$, which is the history of the game up to round t. In other words, $p_t(a)$ itself is a random variable.

[5] We note that the original algorithm in Auer et al. (2002b) was formulated for the gains game. Here we present an improved algorithm for the losses game (Stoltz, 2005, Bubeck, 2010). We refer to home assignment for the difference between the two.

2. Even though $\tilde{\ell}_t^a$-s are not independent, they are unbiased estimates of the true losses. Specifically,

$$\begin{aligned}
\mathbb{E}\left[\tilde{\ell}_t^a\right] &= \mathbb{E}\left[\frac{\ell_t^a \mathbb{1}(A_t = a)}{p_t(a)}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\ell_t^a \mathbb{1}(A_t = a)}{p_t(a)}\middle| A_1, \ldots, A_{t-1}\right]\right] \\
&= \mathbb{E}\left[\frac{\ell_t^a}{p_t(a)}\mathbb{E}\left[\mathbb{1}(A_t = a)|A_1, \ldots, A_{t-1}\right]\right] \\
&= \mathbb{E}\left[\frac{\ell_t^a}{p_t(a)}p_t(a)\right] \\
&= \ell_t^a.
\end{aligned}$$

   The first expectation above is with respect to $A_1, \ldots, A_t$. In the nested expectations, the external expectation is with respect to $A_1, \ldots, A_{t-1}$ and the internal is with respect to A_t. Note that $p_t(a)$ is a random variable depending on $A_1, \ldots, A_{t-1}$, thus after the conditioning on $A_1, \ldots, A_{t-1}$ it is deterministic.

3. Since $\ell_t^a \in [0, 1]$, we have $\tilde{\ell}_t^a \in \left[0, \frac{1}{p_t(a)}\right]$.

4. What is important is that the second moment of $\tilde{\ell}_t^a$-s is by an order of magnitude smaller than the second moment of a general random variable in the corresponding range. This is because the expectation of $\tilde{\ell}_t^a$-s is in the $[0, 1]$ interval. Specifically:

$$\begin{aligned}
\mathbb{E}\left[\left(\tilde{\ell}_t^a\right)^2\right] &= \mathbb{E}\left[\left(\frac{\ell_t^a \mathbb{1}(A_t = a)}{p_t(a)}\right)^2\right] \\
&= \mathbb{E}\left[\frac{\left(\ell_t^a\right)^2 \left(\mathbb{1}(A_t = a)\right)^2}{p_t(a)^2}\right] \\
&= \mathbb{E}\left[\frac{\left(\ell_t^a\right)^2 \mathbb{1}(A_t = a)}{p_t(a)^2}\right] \\
&\leq \mathbb{E}\left[\frac{\mathbb{1}(A_t = a)}{p_t(a)^2}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\frac{\mathbb{1}(A_t = a)}{p_t(a)^2}\middle| A_1, \ldots, A_{t-1}\right]\right] \\
&= \mathbb{E}\left[\frac{1}{p_t(a)^2}\mathbb{E}\left[\mathbb{1}(A_t = a)|A_1, \ldots, A_{t-1}\right]\right] \\
&= \mathbb{E}\left[\frac{1}{p_t(a)}\right],
\end{aligned}$$

   where we have used $\left(\mathbb{1}(A_t = a)\right)^2 = \mathbb{1}(A_t = a)$ and $\left(\ell_t^a\right)^2 \leq 1$ (since $\ell_t^a \in [0, 1]$).

Analysis Now we are ready to present the analysis of the algorithm.

Theorem 5.5. *The expected regret of the EXP3 algorithm with a fixed learning rate η satisfies:*

$$\mathbb{E}\left[R_T\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2}KT.$$

The expected regret is minimized by $\eta = \sqrt{\frac{2 \ln K}{KT}}$, which leads to

$$\mathbb{E}\left[R_T\right] \leq \sqrt{2KT \ln K}.$$

Note that the extra payment for being able to observe just one entry rather than the full column is the multiplicative $\sqrt{K}$ factor in the regret bound.

Proof. The proof of the theorem is based on Lemma 5.2. We note that $\tilde{\ell}_t^a$-s are all non-negative and, thus, by Lemma 5.2 we have:

$$\sum_{t=1}^{T}\sum_a p_t(a)\tilde{\ell}_t^a - \min_a \tilde{L}_T(a) \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_a p_t(a)\left(\tilde{\ell}_t^a\right)^2.$$

By taking expectation of the two sides of the inequality we obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\tilde{\ell}_t^a\right] - \mathbb{E}\left[\min_a \tilde{L}_T(a)\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\left(\tilde{\ell}_t^a\right)^2\right].$$

We note that $\mathbb{E}\left[\min\left[\cdot\right]\right] \leq \min\left[\mathbb{E}\left[\cdot\right]\right]$ and thus:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\tilde{\ell}_t^a\right] - \min_a \mathbb{E}\left[\tilde{L}_T(a)\right] \leq \frac{\ln K}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\left(\tilde{\ell}_t^a\right)^2\right].$$

And now we consider the three expectation terms in this inequality.

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\tilde{\ell}_t^a\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_a \mathbb{E}\left[p_t(a)\tilde{\ell}_t^a \middle| A_1,\ldots,A_{t-1}\right]\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\ell_t^a\right],$$

which is the expected loss of EXP3.

$$\mathbb{E}\left[\tilde{L}_T(a)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\tilde{\ell}_t^a\right] = \sum_{t=1}^{T}\ell_t^a,$$

which is the cumulative loss of row a up to time T. And, finally,

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\left(\tilde{\ell}_t^a\right)^2\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_a \mathbb{E}\left[p_t(a)\left(\tilde{\ell}_t^a\right)^2 \middle| A_1,\ldots,A_{t-1}\right]\right] \leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_a p_t(a)\frac{1}{p_t(a)}\right] = KT.$$

Putting all three together back into the inequality we obtain the first statement of the theorem. And, as before, we find η that minimizes the bound. □

5.5.1 Lower Bound

The lower bound is based on construction of $K+1$ games. In the 0-th game all losses are Bernoulli with bias $1/2$. In the i-th game for $i \in \{1,\ldots,K\}$ all losses are Bernoulli with bias $1/2$ except the losses of the i-th arm, which are Bernoulli with bias $1/2-\varepsilon$ for $\varepsilon = \sqrt{cK/T}$, where c is a properly selected constant. With T/K pulls it is impossible to distinguish between Bernoulli random variable with bias $1/2$ and Bernoulli random variable with bias $1/2-\sqrt{K/T}$, because they induce indistinguishable distributions over sequences of length T/K. As a result, within T pulls the player cannot distinguish between the 0-th game and the i-th games. Therefore, if the adversary picks an i-th game at random the player's regret will on average (with respect to the adversary's and the players choices) be at least $\Omega\left(\varepsilon T\right) = \Omega\left(\sqrt{KT}\right)$. For the details of the proof see Cesa-Bianchi and Lugosi (2006), Bubeck and Cesa-Bianchi (2012).

It is possible to close the $\sqrt{\ln K}$ gap between the upper and the lower bound by modifying the algorithm and improving the upper bound. See Bubeck and Cesa-Bianchi (2012) for details.

5.6 Adversarial Multiarmed Bandits with Expert Advice

Game setting We are, again, working with the same matrix of losses as in prediction with expert advice. But now on every round of the game we get advice of N experts indexed by h in a form of a distribution over the K arms. More formally:
For $t = 1, 2, \ldots$:

1. Observe $q_t^1, \ldots, q_t^N$, where q_t^h is a probability distribution over $\{1,\ldots,K\}$.
2. Pick a row A_t.
3. Observe & suffer $\ell_t^{A_t}$. (ℓ_t^a-s for $a \neq A_t$ remain unobserved)

Performance measure We compare the expected loss of the algorithm to the expected loss of the best expert, where the expectation of the loss of expert h is taken with respect to its advice vector q_h. Specifically:

$$\mathbb{E}[R_T] = \sum_{t=1}^{T}\sum_{a} p_t(a)\ell_t^a - \min_h \sum_{t=1}^{T}\sum_{a} q_t^h(a)\ell_t^a.$$

Algorithm The algorithm is quite similar to the EXP3 algorithm.[6] Note that now $\tilde{L}_t(h)$ tracks cumulative importance-weighted estimate of expert losses instead of individual arm losses.

Algorithm 6 EXP4 (Auer et al., 2002b)

Input: Learning rates $\eta_1 \geq \eta_2 \geq \cdots > 0$
$\forall h : \tilde{L}_0(h) = 0$
for $t = 1, 2, \ldots$ **do**
  $\forall h$: $w_t(h) = \frac{e^{-\eta_t \tilde{L}_{t-1}(h)}}{\sum_{h'} e^{-\eta_t \tilde{L}_{t-1}(h')}}$
  Observe $q_t^1, \ldots, q_t^N$
  $\forall a$: $p_t(a) = \sum_h w_t(h) q_t^h(a)$
  Sample A_t according to p_t and play it
  Observe and suffer $\ell_t^{A_t}$
  Set $\tilde{\ell}_t^a = \frac{\ell_t^a \mathbb{1}(A_t = a)}{p_t(a)} = \begin{cases} \frac{\ell_t^a}{p_t(a)}, & \text{If } A_t = a \\ 0, & \text{otherwise} \end{cases}$
  Set $\tilde{\ell}_t^h = \sum_a q_t(h)\tilde{\ell}_t^a$
  $\forall h$: $\tilde{L}_t(h) = \tilde{L}_{t-1}(h) + \tilde{\ell}_t^h$
end for

Analysis The EXP4 algorithm satisfies the following regret guarantee.

Theorem 5.6. *The expected regret of the EXP4 algorithm with a fixed learning rate η satisfies:*

$$\mathbb{E}[R_T] \leq \frac{\ln N}{\eta} + \frac{\eta}{2} KT.$$

The expected regret is minimized by $\eta = \sqrt{\frac{2 \ln N}{KT}}$, which leads to

$$\mathbb{E}[R_T] \leq \sqrt{2KT \ln N}.$$

Note that $\ln N$ term plays the role of complexity of the class of experts in a very similar way to the complexity terms we saw earlier in supervised learning (specifically, in the uniform union bound).

Proof. The analysis is quite similar to the analysis of the EXP3 algorithm. We note that $\tilde{\ell}_t^h$-s are all non-negative and that w_t is a distribution over $\{1, \ldots, N\}$ defined in the same way as p_t in Lemma 5.2. Thus, by Lemma 5.2 we have:

$$\sum_{t=1}^{T}\sum_{h} w_t(h)\tilde{\ell}_t^h - \min_h \tilde{L}_T(h) \leq \frac{\ln N}{\eta} + \frac{\eta}{2}\sum_{t=1}^{T}\sum_{h} w_t(h)\left(\tilde{\ell}_t^h\right)^2.$$

By taking expectations of the two sides of this expression we obtain:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\tilde{\ell}_t^h\right] - \mathbb{E}\left[\min_h \tilde{L}_T(h)\right] \leq \frac{\ln N}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\left(\tilde{\ell}_t^h\right)^2\right].$$

[6] As with the EXP3 algorithm we present a slightly improved version of the algorithm for the game with losses. The original algorithm was designed for the game with rewards.

As before, $\mathbb{E}\left[\min\left[\cdot\right]\right] \leq \min\left[\mathbb{E}\left[\cdot\right]\right]$ and thus:

$$\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\tilde{\ell}_t^h\right] - \min_h \mathbb{E}\left[\tilde{L}_T(h)\right] \leq \frac{\ln N}{\eta} + \frac{\eta}{2}\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\left(\tilde{\ell}_t^h\right)^2\right].$$

And now we consider the three expectation terms in this inequality.

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\tilde{\ell}_t^h\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\sum_{a} q_t^h(a)\tilde{\ell}_t^a\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\sum_{h} w_t(h)q_t^h(a)\right)\tilde{\ell}_t^a\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a} p_t(a)\tilde{\ell}_t^a\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a} p_t(a)\ell_t^a\right],
\end{aligned}$$

where the first equality is by the definition of $\tilde{\ell}_t^h$ and the last equality is due to unbiasedness of $\tilde{\ell}_t^a$. Thus, the first expectation is the expected loss of EXP4.

$$\mathbb{E}\left[\tilde{L}_T(h)\right] = \mathbb{E}\left[\sum_{t=1}^{T}\tilde{\ell}_t^h\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a} q_t(a)\tilde{\ell}_t^a\right] = \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a} q_t(a)\ell_t^a\right],$$

where we can remove tilde due to unbiasedness of $\tilde{\ell}_t^a$ and we obtain the expected cumulative loss of expert h over T rounds. And, finally,

$$\begin{aligned}
\mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\left(\tilde{\ell}_t^h\right)^2\right] &= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\left(\sum_{a} q_t^h(a)\tilde{\ell}_t^a\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{t=1}^{T}\sum_{h} w_t(h)\sum_{a} q_t^h(a)\left(\tilde{\ell}_t^a\right)^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a}\left(\sum_{h} w_t(h)q_t^h(a)\right)\left(\tilde{\ell}_t^a\right)^2\right] \\
&= \mathbb{E}\left[\sum_{t=1}^{T}\sum_{a} p_t(a)\left(\tilde{\ell}_t^a\right)^2\right] \\
&\leq KT,
\end{aligned}$$

where the first inequality is by Jensen's inequality and convexity of x^2 and the last inequality is along the same lines as the analogous inequality in the analysis of EXP3. By substituting the three expectations back into the inequality we obtain the first statement of the theorem. And, as before, we find η that minimizes the bound. □

5.6.1 Lower Bound

It is possible to show that the regret of adversarial multiarmed bandits with expert advice must be at least $\Omega\left(\sqrt{KT\frac{\ln N}{\ln K}}\right)$. The lower bound is based on construction of $\frac{\ln N}{\ln K}$ independent bandit problems, each according to the construction of the lower bound for multiarmed bandits in Section 5.5.1, and construction of expert advice, so that for every possible selection of best arms for the subproblems there is an expert that recommends that selection. For details of the proof see Agarwal et al. (2012), Seldin and Lugosi (2016). Closing the $\sqrt{\ln K}$ gap between the upper and the lower bound is an open problem.

Appendix A

Set Theory Basics

In this chapter we provide a number of basic definitions and notations from the set theory that are used in the notes.

Countable and Uncountable sets A set is called *countable* if its elements can be counted or, in other words, if every element in a set can be associated with a natural number. For example, the set of integer numbers is countable and the set of rational numbers (ratios of two integers) is also countable. Finite sets are countable as well. A set is called *uncountable* if its elements cannot be enumerated. For example, the set of real numbers $\mathbb{R}$ is uncountable and the set of numbers in a $[0, 1]$ interval is also uncountable.

Relations between sets For two sets A and B we use $A \subseteq B$ to denote that A is a subset of B.

Operations on sets For two sets A and B we use $A \cup B$ to denote the union of A and B; $A \cap B$ the intersection of A and B; and $A \setminus B$ the difference of A and B (the set of elements that are in A, but not in B).

The empty set We use $\varnothing$ or $\{\}$ to denote the empty set.

Disjoint sets Two sets A and B are called *disjoint* if $A \cap B = \varnothing$.

Appendix B

Probability Theory Basics

This chapter provides a number of basic definitions and results from the probability theory. It is partially based on Mitzenmacher and Upfal (2005).

B.1 Axioms of Probability

We start with a definition of a probability space.

Definition B.1 (Probability space)**.** *A probability space is a tuple* $(\Omega, \mathcal{F}, \mathbb{P})$*, where*

- Ω *is a* sample space*, which is the set of all possible outcomes of the random process modeled by the probability space.*
- $\mathcal{F}$ *is a family of sets representing the allowable events, where each set in* $\mathcal{F}$ *is a subset of the sample space* Ω*.*
- $\mathbb{P}$ *is a probability function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *satisfying Definition B.4.*

Elements of Ω are called *simple* or *elementary* events.

Example B.2. For coin flips the sample space is $\Omega = \{H, T\}$, where H stands for "heads" and T for "tails".

In dice rolling the sample space is $\Omega = \{1, 2, 3, 4, 5, 6\}$, where 1,...,6 label the sides of a dice (you should consider them as labels rather than numerical values, we get back to this later in Example B.15).

If we simultaneously flip a coin and roll a dice the sample space is
$\Omega = \{(H, 1), (T, 1), (H, 2), (T, 2), \ldots, (H, 6), (T, 6)\}$.

If Ω is countable (including finite), the probability space is *discrete*. In discrete probability spaces the family $\mathcal{F}$ consists of all subsets of Ω. In particular, $\mathcal{F}$ always includes the empty set $\varnothing$ and the complete sample space Ω. If Ω is uncountably infinite (for example, the real line or the $[0, 1]$ interval) a proper definition of $\mathcal{F}$ requires concepts from the measure theory, which go beyond the scope of these notes.

Example B.3. In the coin flipping experiment $\mathcal{F} = \{\varnothing, \{H\}, \{T\}, \{H, T\}\}$.

Definition B.4 (Probability Axioms)**.** *A probability function is any function* $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ *that satisfies the following conditions*

1. *For any event* $E \in \mathcal{F}$*,* $0 \leq \mathbb{P}(E) \leq 1$*.*
2. $\mathbb{P}(\Omega) = 1$*.*
3. *For any finite or countably infinite sequence of mutually disjoint events* $E_1, E_2, \ldots$

$$\mathbb{P}\left(\bigcup_{i \geq 1} E_i\right) = \sum_{i \geq 1} \mathbb{P}(E_i).$$

We now consider a number of basic properties of probabilities.

Lemma B.5 (Monotonicity)**.** *Let A and B be two events, such that $A \subseteq B$. Then*

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof. We have that $B = A \cup (B \setminus A)$ and the events A and $B \setminus A$ are disjoint. Thus,

$$\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \setminus A) \geq \mathbb{P}(A),$$

where the equality is by the third axiom of probabilities and the inequality is by the first axiom of probabilities, since $\mathbb{P}(B \setminus A) \geq 0$. □

The next simple, but very important result is known as the *union bound*.

Lemma B.6 (The union bound)**.** *For any finite or countably infinite sequence of events $E_1, E_2, \ldots$,*

$$\mathbb{P}\left(\bigcup_{i\geq 1} E_i\right) \leq \sum_{i\geq 1} \mathbb{P}(E_i).$$

Proof. We have

$$\bigcup_{i\geq 1} E_i = E_1 \cup (E_2 \setminus E_1) \cup (E_3 \setminus (E_1 \cup E_2)) \cup \cdots = \bigcup_{i\geq 1} F_i,$$

where the events $F_i = E_i \setminus \bigcup_{j=1}^{i-1} E_j$ are disjoint, $F_i \subseteq E_i$, and $\bigcup_{i\geq 1} F_i = \bigcup_{i\geq 1} E_i$. Therefore,

$$\mathbb{P}\left(\bigcup_{i\geq 1} E_i\right) = \mathbb{P}\left(\bigcup_{i\geq 1} F_i\right) = \mathbb{P}\left(\bigcup_{i\geq 1} F_i\right) = \sum_{i\geq 1} \mathbb{P}(F_i) \leq \sum_{i\geq 1} \mathbb{P}(E_i),$$

where the second equality is by the third axiom of probabilities and the inequality is by monotonicity of the probability (Lemma B.5). □

Example B.7. Let $E_1 = \{1, 3, 5\}$ be the event that the outcome of a dice roll is odd and $E_2 = \{1, 2, 3\}$ be the event that the outcome is at most 3. Then $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(1, 2, 3, 5) \leq \mathbb{P}(E_1) + \mathbb{P}(E_2)$. Note that this is true irrespective of the choice of the probability measure $\mathbb{P}$. In particular, this is true irrespective of whether the dice is fair or not.

Definition B.8 (Independence)**.** *Two events A and B are called* independent *if and only if*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B).$$

Definition B.9 (Pairwise independence)**.** *Events $E_1, \ldots, E_n$ are called* pairwise independent *if and only if for any pair i, j*

$$\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i)\mathbb{P}(E_j).$$

Definition B.10 (Mutual independence)**.** *Events $E_1, \ldots, E_n$ are called* mutually independent *if and only if for any subset of indices $I \subseteq \{1, \ldots, n\}$*

$$\mathbb{P}\left(\bigcap_{i\in I} E_i\right) = \prod_{i\in I} \mathbb{P}(E_i).$$

Note that pairwise independence does not imply mutual independence. Take the following example: assume we roll a fair tetrahedron (a three-dimensional object with four faces) with faces colored in red, blue, green, and the fourth face colored in all three colors, red, blue, and green. Let E_1 be the event that we observe red color, E_2 be the even that we observe blue color, and E_3 be the event that we observe green color. Then for all i we have $\mathbb{P}(E_i) = \frac{1}{2}$ and for any pair $i \neq j$ we have $\mathbb{P}(E_i \cap E_j) = \frac{1}{4} = \mathbb{P}(E_i)\mathbb{P}(E_j)$. However, $\mathbb{P}(E_1 \cap E_2 \cap E_3) = \frac{1}{4} \neq \mathbb{P}(E_1)\mathbb{P}(E_2)\mathbb{P}(E_3)$ and, thus, the events are pairwise independent, but not mutually independent. If we say that events $E_1, \ldots, E_n$ are independent without further specifications we imply mutual independence.

Definition B.11 (Conditional probability)**.** *The* conditional probability *that event A occurs given that event B occurs is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

The conditional probability is well-defined only if $\mathbb{P}(B) > 0$.

By the definition we have that $\mathbb{P}(A \cap B) = \mathbb{P}(B)\mathbb{P}(A|B) = \mathbb{P}(A)\mathbb{P}(B|A)$.

Example B.12. For a fair dice let $A = \{1, 6\}$ and $B = \{1, 2, 3, 4\}$. Then

$$\begin{aligned} \mathbb{P}(A) &= \frac{1}{3}, \\ \mathbb{P}(B) &= \frac{2}{3}, \\ A \cap B &= \{1\}, \\ \mathbb{P}(A \cap B) &= \frac{1}{6}, \\ \mathbb{P}(A|B) &= \frac{\frac{1}{6}}{\frac{2}{3}} = \frac{1}{4}. \end{aligned}$$

Lemma B.13 (The law of total probability)**.** *Let $E_1, E_2, \ldots, E_n$ be mutually disjoint events, such that $\bigcup_{i=1}^{n} E_n = \Omega$. Then*

$$\mathbb{P}(A) = \sum_{i=1}^{n} \mathbb{P}(A \cap E_i) = \sum_{i=1}^{n} \mathbb{P}(A|E_i)\mathbb{P}(E_i).$$

Proof. Since the E_i-s are disjoint and cover the entire space it follows that $A = \bigcup_{i=1}^{n}(A \cap E_i)$ and the events $A \cap E_i$ are mutually disjoint. Therefore,

$$\mathbb{P}(A) = \mathbb{P}\left(\bigcup_{i=1}^{n}(A \cap E_i)\right) = \sum_{i=1}^{n} \mathbb{P}(A \cap E_i) = \sum_{i=1}^{n} \mathbb{P}(A|E_i)\mathbb{P}(E_i).$$

□

B.2 Discrete Random Variables

We now define another basic concept in probability theory, a *random variable*.

Definition B.14. *A* random variable *X on a sample space Ω is a real-valued function on Ω, that is $X : \Omega \to \mathbb{R}$. A* discrete *random variable is a random variable that takes on only a finite or countably infinite number of values.*

Example B.15. For a coin we can define a random variable X, such that $X(H) = 1$ and $X(T) = 0$. We can also define another random variable Y, such that $Y(H) = 1$ and $Y(T) = -1$.

For a dice we can define a random variable X, such that $X(1) = 1, X(2) = 2, X(3) = 3, X(4) = 4, X(5) = 5, X(6) = 6$. We can also define a random variable Y, such that $Y(1) = 3, Y(2) = 2.4, Y(3) = -6, Y(4) = 8, Y(5) = 8, Y(6) = 0$. This example emphasizes the difference between labeling of events and assignment of numerical values to events. Note that the random variable Y does not distinguish between faces 4 and 5 of the dice, even though they are separate events in the probability space.

Functions of random variables are also random variables. In the last example, a random variable $Z = X^2$ takes values $Z(1) = 1, Z(2) = 4, Z(3) = 9, \ldots, Z(6) = 36$.

Definition B.16 (Independence of random variables)**.** *Two random variables X and Y are* independent *if and only if*

$$\mathbb{P}((X = x) \cap (Y = y)) = \mathbb{P}(X = x)\mathbb{P}(Y = y).$$

for all values x and y.

Definition B.17 (Pairwise independence)**.** *Random variables $X_1, \ldots, X_n$ are* pairwise independent *if and only if for any pair i, j and any values x_i, x_j*

$$\mathbb{P}((X_i = x_i) \cap (X_j = x_j)) = \mathbb{P}(X_i = x_i)\mathbb{P}(X_j = x_j).$$

Definition B.18 (Mutual independence)**.** *Random variables $X_1, \ldots, X_n$ are* mutually independent *if for any subset of indices $I \subseteq \{1, \ldots, n\}$ and any values x_i, $i \in I$*

$$\mathbb{P}\left(\bigcap_{i \in I}(X_i = x_i)\right) = \prod_{i \in I}\mathbb{P}(X_i = x_i).$$

Similar to the example given earlier, pairwise independence of random variables does not imply their mutual independence. If we say that random variables are independent without further specifications we imply mutual independence.

B.3 Expectation

Expectation is the most basic characteristic of a random variable.

Definition B.19 (Expectation)**.** *Let X be a discrete random variable and let $\mathcal{X}$ be the set of all possible values that it can take. The* expectation *of X, denoted by $\mathbb{E}[X]$, is given by*

$$\mathbb{E}[X] = \sum_{x \in \mathcal{X}} x\mathbb{P}(X = x).$$

The expectation is finite if $\sum_{x \in \mathcal{X}} |x|\mathbb{P}(X = x)$ converges; otherwise the expectation is unbounded.

Example B.20. For a fair dice with faces numbered 1 to 6 let $X(i) = i$ (the i-th face gets value i). Then

$$\mathbb{E}[X] = \sum_{i=1}^{6} i\frac{1}{6} = \frac{7}{2}.$$

Take another random variable $Z = X^2$ then

$$\mathbb{E}[Z] = \mathbb{E}\left[X^2\right] = \sum_{i=1}^{6} i^2\frac{1}{6} = \frac{91}{6}.$$

Expectation satisfies a number of important properties (these properties also hold for continuous random variables). We leave a proof of these properties as an exercise.

Lemma B.21 (Multiplication by a constant)**.** *For any constant c*

$$\mathbb{E}[cX] = c\mathbb{E}[X].$$

Theorem B.22 (Linearity)**.** *For any pair of random variables X and Y, not necessarily independent,*

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Theorem B.23. *If X and Y are independent random variables, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y].$$

We emphasize that in contrast with Theorem B.22, this property does not hold in the general case (if X and Y are not independent).

B.4 Variance

Variance is the second most basic characteristic of a random variable.

Definition B.24 (Variance)**.** *The* variance *of a random variable X (discrete or continuous), denoted by* $\operatorname{Var}[X]$*, is defined by*

$$\operatorname{Var}[X] = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$

We invite the reader to prove that $\mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2$.

Example B.25. For a fair dice with faces numbered 1 to 6 let $X(i) = i$ (the i-th face gets value i). Then

$$\operatorname{Var}[X] = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2 = \frac{91}{6} - \frac{49}{4} = \frac{35}{12}.$$

Theorem B.26. *If $X_1, \ldots, X_n$ are independent random variables then*

$$\operatorname{Var}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \operatorname{Var}[X_i].$$

The proof is based on Theorem B.23 and the result does not necessarily hold when X_i-s are not independent. We leave the proof as an exercise.

B.5 The Bernoulli and Binomial Random Variables

Two most basic discrete random variables are Bernoulli and binomial.

Definition B.27 (Bernoulli random variable)**.** *A random variable X taking values $\{0, 1\}$ is called a* Bernoulli random variable*. The parameter $p = \mathbb{P}(X = 1)$ is called the* bias *of X.*

Bernoulli random variable has the following property (which does not hold in general):

$$\mathbb{E}[X] = 0 \cdot (1 - p) + 1 \cdot p = p = \mathbb{P}(X = 1).$$

Definition B.28 (Binomial random variable)**.** *A* binomial random variable *Y with parameters n and p, denoted by $B(n, p)$, is defined by the following probability distribution on $k \in \{0, 1, \ldots, n\}$:*

$$\mathbb{P}(Y = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$

Binomial random variable can be represented as a sum of independent identically distributed Bernoulli random variables.

Lemma B.29. *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with bias p. Then $Y = \sum_{i=1}^{n} X_i$ is a binomial random variable with parameters n and p.*

A proof of this lemma is left as an exercise to the reader.

B.6 Jensen's Inequality

Jensen's inequality is one of the most basic in probability theory.

Theorem B.30 (Jensen's inequality)**.** *If f is a convex function and X is a random variable, then*

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X]).$$

For a proof see, for example, Mitzenmacher and Upfal (2005) or Cover and Thomas (2006).

Appendix C

Linear Algebra

We revisit a number of basic concepts from linear algebra. This is only a brief revision of the main concepts that we are using in the lecture notes. For more details, please, refer to Strang (2009) or some other textbook on linear algebra.

We start with reminding that two vectors $\mathbf{u}$ and $\mathbf{v}$ are perpendicular, $\mathbf{u} \perp \mathbf{v}$, if and only if their inner product $\mathbf{u}^T\mathbf{v} = 0$.

Matrix A matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ takes vectors in $\mathbb{R}^d$ and maps them into $\mathbb{R}^n$. There are two fundamental subspaces associated with a matrix $\mathbf{X}$. The *image* of $\mathbf{X}$, denoted $Im(\mathbf{X}) \subseteq \mathbb{R}^n$, is the space of all vectors $\mathbf{v} \in \mathbb{R}^n$ that can be obtained through multiplication of $\mathbf{X}$ with a vector $\mathbf{w}$. The image $Im(\mathbf{X})$ is a linear subspace of $\mathbb{R}^n$ and it is also called a *column space* of $\mathbf{X}$. The second subspace is the *nullspace* of $\mathbf{X}$, denoted $Null(\mathbf{X}) \subseteq \mathbb{R}^d$, which is the space of all vectors $\mathbf{w}$ for which $\mathbf{X}\mathbf{w} = 0$. The nullspace is a linear subspace of $\mathbb{R}^d$. The subspaces are illustrated in Figure C.1.

Matrix transpose Matrix transpose $\mathbf{X}^T$ takes vectors in $\mathbb{R}^n$ and maps them into $\mathbb{R}^d$. The corresponding subspaces are $Im(\mathbf{X}^T)$, the *row space* of $\mathbf{X}$, and $Null(\mathbf{X}^T)$.

Orthogonality of the fundamental subspaces $Im(\mathbf{X}) \perp Null(\mathbf{X}^T)$ and $Im(\mathbf{X}^T) \perp Null(\mathbf{X})$ There is an important and extremely beautiful relation between the four fundamental subspaces associated with a matrix $\mathbf{X}$ and its transpose. Namely, the image of $\mathbf{X}$ is orthogonal to the nullspace of $\mathbf{X}^T$ and the image of $\mathbf{X}^T$ is orthogonal to the nullspace of $\mathbf{X}$. It means that if we take any two vectors $\mathbf{u} \in Im(\mathbf{X})$ and $\mathbf{v} \in Null(\mathbf{X}^T)$ then $\mathbf{u}^T\mathbf{v} = 0$ (and the same for the second pair of subspaces). The proof of this fact is short and elegant. Any vector in $\mathbf{u} \in Im(\mathbf{X})$ can be represented as a linear combination of the rows of $\mathbf{X}$, meaning that $\mathbf{u} = \mathbf{X}\mathbf{z}$. At the same time, by definition of a nullspace, if $\mathbf{v} \in Null(\mathbf{X}^T)$ then $\mathbf{X}^T\mathbf{v} = 0$. By putting these two facts together we obtain:

$$\mathbf{u}^T\mathbf{v} = (\mathbf{X}\mathbf{z})^T\mathbf{v} = \mathbf{z}^T\mathbf{X}^T\mathbf{v} = \mathbf{z}^T(\mathbf{X}^T\mathbf{v}) = 0.$$

Complete relation between $Im(\mathbf{X})$, $Im(\mathbf{X}^T)$, $Null(\mathbf{X})$, and $Null(\mathbf{X}^T)$ Not only the pairs $Im(\mathbf{X})$ with $Null(\mathbf{X}^T)$ and $Im(\mathbf{X}^T)$ with $Null(\mathbf{X})$ are orthogonal, they also complement each other. Let $dim(\mathbf{A})$ denote dimension of a matrix $\mathbf{A}$. The dimension is equal to the number of independent columns, which is equal to the number of independent rows (this fact can be shown by bringing $\mathbf{A}$ to a diagonal form). Then we have the following relations:

1. $dim(Im(\mathbf{X})) = dim(Im(\mathbf{X}^T)) = dim(\mathbf{X})$.
2. $dim(Null(\mathbf{X})) = d - dim(Im(\mathbf{X}^T))$ and $dim(Null(\mathbf{X}^T)) = n - dim(Im(\mathbf{X}))$.
3. $Im(\mathbf{X}) \perp Null(\mathbf{X}^T)$ and $Im(\mathbf{X}^T) \perp Null(X)$.

Together these properties mean that a combination of bases for $Im(\mathbf{X}^T)$ and $Null(\mathbf{X})$ makes a basis for $\mathbb{R}^d$ and a combination of bases for $Im(\mathbf{X})$ and $Null(\mathbf{X}^T)$ make a basis for $\mathbb{R}^n$. It means that any vector $\mathbf{v} \in \mathbb{R}^d$ can be represented as $\mathbf{v} = \mathbf{v}_\star + \mathbf{v}_0$, where $\mathbf{v}_\star \in Im(\mathbf{X}^T)$ belongs to the row space of $\mathbf{X}$ and $v_0 \in Null(\mathbf{X})$ belongs to the nullspace of $\mathbf{X}$.

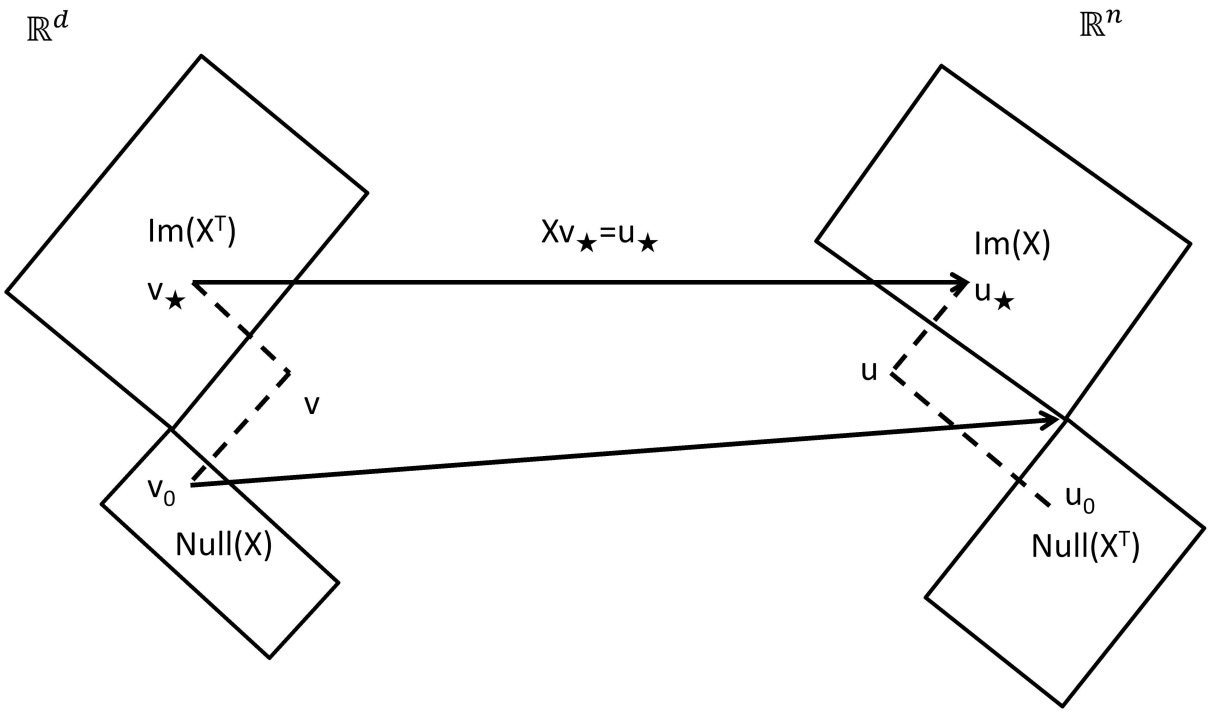


Figure C.1: **The four fundamental subspaces of a matrix X.** There is a right angle between $Im(\mathbf{X})$ and $Null(\mathbf{X}^T)$, as well as between $Im(\mathbf{X}^T)$ and $Null(\mathbf{X})$.

The mapping between $Im(\mathbf{X}^T)$ and $Im(\mathbf{X})$ is one-to-one and, thus, invertible Every vector $\mathbf{u}$ in the column space comes from one and only one vector in the row space $\mathbf{v}$. The proof of this fact is also simple. Assume that $\mathbf{u} = \mathbf{X}\mathbf{v} = \mathbf{X}\mathbf{v}'$ for two vectors $\mathbf{v}, \mathbf{v}' \in Im(\mathbf{X}^T)$. Then $\mathbf{X}(\mathbf{v} - \mathbf{v}') = 0$ and the vector $\mathbf{v} - \mathbf{v}' \in Null(\mathbf{X})$. But $Null(\mathbf{X})$ is perpendicular to $Im(\mathbf{X}^T)$, which means that $\mathbf{v} - \mathbf{v}'$ is orthogonal to itself and, therefore, must be the zero vector.

$\mathbf{X}^T\mathbf{X}$ is invertible if and only if $\mathbf{X}$ has linearly independent columns $(\mathbf{X}^T\mathbf{X})^{-1}$ is a very important matrix. We show that $\mathbf{X}^T\mathbf{X}$ is invertible if and only if $\mathbf{X}$ has linearly independent columns, meaning that $dim(X) = d$. We show this by proving that $\mathbf{X}$ and $\mathbf{X}^T\mathbf{X}$ have the same nullspace. Let $\mathbf{v} \in Null(\mathbf{X})$, then $\mathbf{X}\mathbf{v} = 0$ and, therefore, $\mathbf{X}^T\mathbf{X}\mathbf{v} = 0$ and $\mathbf{v} \in Null(\mathbf{X}^T\mathbf{X})$. In the other direction, let $\mathbf{v} \in Null(\mathbf{X}^T\mathbf{X})$. Then $\mathbf{X}^T\mathbf{X}\mathbf{v} = 0$ and we have:

$$\|\mathbf{X}\mathbf{v}\|^2 = (\mathbf{X}\mathbf{v})^T(\mathbf{X}\mathbf{v}) = \mathbf{v}^T\mathbf{X}^T\mathbf{X}\mathbf{v} = \mathbf{v}^T(\mathbf{X}^T\mathbf{X}\mathbf{v}) = 0.$$

Since $\|\mathbf{X}\mathbf{v}\|^2 = 0$ if and only if $\mathbf{X}\mathbf{v} = 0$, we have $\mathbf{v} \in Null(\mathbf{X})$.

$\mathbf{X}^T\mathbf{X}$ is an $d \times d$ square matrix, therefore $dim(\mathbf{X}^T\mathbf{X}) = d - dim(Null(\mathbf{X}^T\mathbf{X}) = d - dim(Null(\mathbf{X}))$ and matrix $\mathbf{X}^T\mathbf{X}$ is invertible if and only if the dimension of the nullspace of $\mathbf{X}$ is zero, meaning that $\mathbf{X}$ has linearly independent columns. (Note that unless $n = d$, $\mathbf{X}$ itself is a rectangular matrix and that inverses are not defined for rectangular matrices.)

Projection onto a line A line in direction $\mathbf{u}$ is described by $\alpha\mathbf{u}$ for $\alpha \in \mathbb{R}$. Projection of vector $\mathbf{v}$ onto vector $\mathbf{u}$ means that we are looking for a projection vector $\mathbf{p} = \alpha\mathbf{u}$, such that the remainder $\mathbf{v} - \mathbf{p}$ is orthogonal to the projection. So we have:

$$\begin{aligned}(\mathbf{v} - \alpha\mathbf{u})^T\alpha\mathbf{u} &= 0,\\ \alpha\mathbf{v}^T\mathbf{u} &= \alpha^2\mathbf{u}^T\mathbf{u},\\ \alpha &= \frac{\mathbf{v}^T\mathbf{u}}{\mathbf{u}^T\mathbf{u}} = \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}.\end{aligned}$$

Thus, the projection $\mathbf{p} = \alpha\mathbf{u} = \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\mathbf{u}$. Note that $\frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}$ is a scalar, thus

$$\mathbf{p} = \frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}}\mathbf{u} = \mathbf{u}\frac{\mathbf{u}^T\mathbf{v}}{\mathbf{u}^T\mathbf{u}} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}\mathbf{v}.$$

The matrix $\mathbf{P} = \frac{\mathbf{u}\mathbf{u}^T}{\mathbf{u}^T\mathbf{u}}$ is a *projection matrix*. For any vector $\mathbf{v}$ the matrix $\mathbf{P}$ projects $\mathbf{v}$ onto u.

Projection onto a subspace A subspace can be described by a set of linear combinations $\mathbf{Az}$, where the columns of matrix $\mathbf{A}$ span the subspace. Projection of a vector $\mathbf{v}$ onto a subspace described by $\mathbf{A}$ means that we are looking for a projection $\mathbf{p} = \mathbf{Az}$, such that the remainder $\mathbf{v} - \mathbf{p}$ is perpendicular to the projection. The projection $\mathbf{p} = \mathbf{Az}$ belongs to the image of $\mathbf{A}$, $Im(\mathbf{A})$. Thus, the remainder must be in the nullspace of $\mathbf{A}^T$, meaning that $\mathbf{A}^T(\mathbf{v} - \mathbf{p}) = 0$. Assuming that the columns of A are independent, we have:

$$\begin{aligned}\mathbf{A}^T(\mathbf{v} - \mathbf{Az}) &= 0,\\ \mathbf{A}^T\mathbf{v} &= \mathbf{A}^T\mathbf{Az},\\ \mathbf{z} &= (\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{v},\end{aligned}$$

where we used independence of the columns of $\mathbf{A}$ in the last step to invert $\mathbf{A}^T\mathbf{A}$. The projection is $\mathbf{p} = \mathbf{Az} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T\mathbf{v}$ and the projection matrix is $\mathbf{P} = \mathbf{A}(\mathbf{A}^T\mathbf{A})^{-1}\mathbf{A}^T$. The projection matrix $\mathbf{P}$ maps any vector $\mathbf{v}$ onto space spanned by the columns of $\mathbf{A}$, $Im(\mathbf{A})$. Note how $(\mathbf{A}^T\mathbf{A})^{-1}$ plays the role of $\frac{1}{\mathbf{u}^T\mathbf{u}}$ in projection onto a line.

Projection matrices Projection matrices satisfy a number of interesting properties:

1. If $\mathbf{P}$ is a projection matrix then $\mathbf{P}^2 = \mathbf{P}$ (the second projection does not change the vector).
2. If $\mathbf{P}$ is a projection matrix projecting onto a subspace described by $\mathbf{A}$ then $\mathbf{I}-\mathbf{P}$ is also a projection matrix. It projects onto a subspace that is perpendicular to the subspace described by $\mathbf{A}$.

Appendix D

Calculus

We revisit some basic concepts from calculus.

D.1 Gradients

Gradients are vectors of partial derivatives. For a vector $\mathbf{x} = \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix}$ and a function $f(\mathbf{x})$ the gradient of f is defined as

$$\nabla f(\bar{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}.$$

Gradient of a multivariate quadratic function $f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$

Let A be a matrix with entries a_{ij}. Then

$$f(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = (x_1, \cdots, x_d) \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} = (x_1, \cdots, x_d) \begin{pmatrix} \sum_{j=1}^d a_{1j}x_j \\ \vdots \\ \sum_{j=1}^d a_{dj}x_j \end{pmatrix} = \sum_{i=1}^d \sum_{j=1}^d a_{ij}x_i x_j.$$

The partial derivative $\frac{\partial f}{\partial x_k}$ then becomes:

$$\frac{\partial f}{\partial x_k} = \frac{\partial \left(\sum_{i=1}^d \sum_{j=1}^d a_{ij}x_i x_j \right)}{\partial x_k} = \sum_{j=1}^d a_{kj}x_j + \sum_{i=1}^d a_{ik}x_i,$$

where the first sum corresponds to the first element in the product $x_i x_j$ being x_k and the second sum corresponds to the second element in the product $x_i x_j$ being x_k. Putting all the derivatives together we obtain:

$$\begin{aligned} \nabla f(\mathbf{x}) &= \begin{pmatrix} \sum_{j=1}^d a_{1j}x_j + \sum_{i=1}^d a_{i1}x_i \\ \vdots \\ \sum_{j=1}^d a_{d1}x_j + \sum_{i=1}^d a_{id}x_i \end{pmatrix} \\ &= \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_d \end{pmatrix} + \left((x_1, \cdots, x_d) \begin{pmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & & \vdots \\ a_{d1} & \cdots & a_{dd} \end{pmatrix} \right)^T \\ &= A\mathbf{x} + A^T\mathbf{x} \\ &= (A + A^T)\mathbf{x}. \end{aligned}$$

A matrix A is called *symmetric* if $A^T = A$. For a symmetric matrix we have $\nabla f(\mathbf{x}) = 2A\mathbf{x}$ and for a general matrix we have $\nabla f(\mathbf{x}) = (A + A^T)\mathbf{x}$. Note the similarity and dissimilarity with the derivative of a univariate quadratic function $f(x) = ax^2$, which is $f'(x) = 2ax$.

Gradient of a linear function $f(\mathbf{x}) = \bar{b}^T\mathbf{x}$

Let $\bar{b} = \begin{pmatrix} b_1 \\ \vdots \\ b_d \end{pmatrix}$ be a vector and let $f(\mathbf{x}) = \bar{b}^T\mathbf{x} = \sum_{i=1}^{d} b_i x_i$. We leave it as an exercise to compute the gradient $\nabla f(\mathbf{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_d} \end{pmatrix}$.

Bibliography

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data*. AMLbook, 2012.

Yaser S. Abu-Mostafa, Malik Magdon-Ismail, and Hsuan-Tien Lin. *Learning from data. Dynamic E-Chapters*. AMLbook, 2015.

Alekh Agarwal, Miroslav Dudík, Satyen Kale, John Langford, and Robert E. Schapire. Contextual bandit learning with predictable rewards. In *Proceedings on the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.

Peter Auer, Nicolò Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 2002a.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal of Computing*, 32(1), 2002b.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration Inequalities A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Sébastien Bubeck. *Bandits Games and Clustering Foundations*. PhD thesis, Université Lille, 2010.

Sébastien Bubeck and Nicolò Cesa-Bianchi. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends in Machine Learning*, 5, 2012.

Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley Series in Telecommunications and Signal Processing, 2nd edition, 2006.

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand. PAC-Bayesian learning of linear classifiers. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2009.

Pascal Germain, Alexandre Lacasse, François Laviolette, Mario Marchand, and Jean-Francis Roy. Risk bounds for the majority vote: From a PAC-Bayesian analysis to a learning algorithm. *Journal of Machine Learning Research*, 16, 2015.

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Tze Leung Lai and Herbert Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6, 1985.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 6, 2005.

Nick Littlestone and Manfred K. Warmuth. The weighted majority algorithm. *Information and Computation*, 108, 1994.

Katalin Marton. A measure concentration inequality for contracting Markov chains. *Geometric and Functional Analysis*, 6(3), 1996.

Katalin Marton. A measure concentration inequality for contracting Markov chains Erratum. *Geometric and Functional Analysis*, 7(3), 1997.

Andrés R. Masegosa, Stephan S. Lorenzen, Christian Igel, and Yevgeny Seldin. Second order PAC-Bayesian bounds for the weighted majority vote. Technical report, `https://arxiv.org/abs/2007.13532`, 2020.

Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.

David McAllester. PAC-Bayesian stochastic model selection. *Machine Learning*, 51, 2003.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press, 2005.

Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, 1952.

Paul-Marie Samson. Concentration of measure inequalities for markov chains and ϕ-mixing processes. *The Annals of Probability*, 28(1), 2000.

Matthias Seeger. PAC-Bayesian generalization error bounds for Gaussian process classification. *Journal of Machine Learning Research*, 3, 2002.

Yevgeny Seldin. The space of online learning problems. ECML-PKDD Tutorial. `https://sites.google.com/site/spaceofonlinelearningproblems/`, 2015.

Yevgeny Seldin and Gábor Lugosi. A lower bound for multi-armed bandits with expert advice. In *Proceedings of the European Workshop on Reinforcement Learning (EWRL)*, 2016.

Yevgeny Seldin, François Laviolette, Nicolò Cesa-Bianchi, John Shawe-Taylor, and Peter Auer. PAC-Bayesian inequalities for martingales. *IEEE Transactions on Information Theory*, 58, 2012.

Gilles Stoltz. *Incomplete Information and Internal Regret in Prediction of Individual Sequences*. PhD thesis, Université Paris-Sud, 2005.

Gilbert Strang. *Introduction to linear algebra*. Wellesley-Cambridge Press, 4^{th} edition, 2009.

Niklas Thiemann, Christian Igel, Olivier Wintenberger, and Yevgeny Seldin. A strongly quasiconvex PAC-Bayesian bound. In *Proceedings of the International Conference on Algorithmic Learning Theory (ALT)*, 2017.

Ilya Tolstikhin and Yevgeny Seldin. PAC-Bayes-Empirical-Bernstein inequality. In *Advances in Neural Information Processing Systems (NIPS)*, 2013.

Vladimir Vovk. Aggregating strategies. In *Proceedings of the Conference on Learning Theory (COLT)*, 1990.