Machine Learning Lecture Notes

Yevgeny Seldin

September 22, 2015

Contents

Chapter 1

Concentration of Measure Inequalities

1.1 Introduction

Machine Learning is (to a significant degree) about using past knowledge to make predictions about the future. Making successful predictions about the future based on past knowledge is possible only when the future is in some way similar to the past and up to the degree that it is similar to the past. One of the simplest and most commonly used ways to model the similarity between past and future are i.i.d. (independent identically distributed) processes. We assume that our past data are independent identically distributed samples from an unknown, but fixed distribution, and that the future observations will be sampled independently from the same distribution. Learning the parameters of this distribution (expectation, variance, etc.) based on past observations then allows to make better predictions about future observations. Thus, concentration of random variables around their expected values ("concentration of measure") plays the central role in theoretical foundations of learning theory. We start with presenting a number of basic results about the concentration of measure phenomenon and then demonstrate applications of these results to learning.

1.2 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 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. By definition of $\mathbb{E}[X]$ we have:

$$\mathbb{E}[X] = \int_0^\infty xp(x)dx = \int_0^\varepsilon xp(x)dx + \int_\varepsilon^\infty xp(x)dx \geq \int_\varepsilon^\infty xp(x)dx \geq \varepsilon \int_\varepsilon^\infty p(x)dx = \varepsilon \mathbb{P}\{X \geq \varepsilon\}.$$

By dividing both sides by ε we obtain the inequality. □

1.3 Hoeffding's Inequality

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

Theorem 2 (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 $\mathbb{P}\{a_i \leq X_i \leq b_i\} = 1$. Then for every $\varepsilon > 0$:*

$$\mathbb{P}\left\{\sum_{i=1}^n X_i - \mathbb{E}\left[\sum_{i=1}^n X_i\right] > \varepsilon\right\} \leq e^{-2\varepsilon^2 / \sum_{i=1}^n (b_i - a_i)^2} \tag{1.1}$$

and

$$\mathbb{P}\left\{\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right] < -\varepsilon\right\} \leq e^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}. \tag{1.2}$$

By taking a union bound of the events in (1.1) and (1.2) we obtain the following corollary.

Corollary 3. *Under the assumptions of Theorem 2:*

$$\mathbb{P}\left\{\left|\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right| > \varepsilon\right\} \leq 2e^{-2\varepsilon^2/\sum_{i=1}^{n}(b_i-a_i)^2}. \tag{1.3}$$

Equations (1.1) and (1.2) are known as "one-sided Hoeffding's inequalities" and (1.3) 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 4. *Let $X_1, \ldots, X_n$ be independent random variables, such that $\mathbb{P}\{X_i \in [0,1]\} = 1$ and $\mathbb{E}[X_i] = \mu$ for all i, then:*

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} X_i - \mu > \varepsilon\right\} \leq e^{-2n\varepsilon^2} \tag{1.4}$$

and

$$\mathbb{P}\left\{\mu - \frac{1}{n}\sum_{i=1}^{n} X_i > \varepsilon\right\} \leq e^{-2n\varepsilon^2}. \tag{1.5}$$

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

Lemma 5 (Hoeffding's Lemma)**.** *Let X be a random variable, such that $\mathbb{P}\{X \in [a,b]\} = 1$. Then for any $\lambda \in \mathbb{R}$:*

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

See a separate handout for a proof of the lemma.

Proof of Theorem 2. We prove the first inequality in Theorem 2. 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:

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

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

$$\begin{aligned}\mathbb{E}\left[e^{\lambda\left(\sum_{i=1}^{n} X_i - \mathbb{E}\left[\sum_{i=1}^{n} X_i\right]\right)}\right] &= \mathbb{E}\left[e^{\left(\sum_{i=1}^{n} \lambda(X_i - \mathbb{E}[X_i])\right)}\right] \\ &= \mathbb{E}\left[\prod_{i=1}^{n} e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \\ &= \prod_{i=1}^{n} \mathbb{E}\left[e^{\lambda(X_i - \mathbb{E}[X_i])}\right] \qquad (1.6) \\ &\leq \prod_{i=1}^{n} e^{\lambda^2(b_i-a_i)^2/8} \qquad (1.7) \\ &= e^{(\lambda^2/8)\sum_{i=1}^{n}(b_i-a_i)^2},\end{aligned}$$

where (1.6) holds since $X_1, \ldots, X_n$ are independent and (1.7) 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 be 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:

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

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. □

1.4 kl Inequality

In this section we derive another concentration inequality, which in certain situations can be much tighter than Hoeffding's inequality. It arises from a close inspection of binomial coefficients and is closely related to the method of types in information theory (Cover and Thomas, 2006, Chapter 11). We start with some definitions.

Definition 6 (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 will have a special interest in Bernoulli random variables.

Definition 7 (Bernoulli random variable)**.** *X is a Bernoulli random variable with bias p if X accepts values in $\{0,1\}$ and $\mathbb{P}\{X=0\} = 1-p$ and $\mathbb{P}\{X=1\} = p$.*

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

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

We specialize the definition of entropy to Bernoulli random variables.

Definition 8 (Binary entropy)**.** *Let p be a bias of Bernoulli random variable X. We define the entropy of p as*

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

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 9.

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

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

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

$$\sum_{i=0}^{n} \binom{n}{i} p^i (1-p)^{n-i} = 1. \tag{1.8}$$

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 (1.8). 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 lower bound. □

Lemma 9 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 ($n-k$).

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

Definition 10 (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 11 (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 12. Let $X_1, \ldots, X_n$ 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 $\frac{1}{n}\sum_{i=1}^n X_i \in \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \frac{n}{n}\right\}$.) Then by Lemma 9:

$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^n X_i = \frac{k}{n}\right\} = \binom{n}{k} p^k (1-p)^{n-k} \leq e^{n\,\mathrm{H}\left(\frac{k}{n}\right)} e^{n\left(\frac{k}{n}\ln p + \frac{n-k}{n}\ln(1-p)\right)} = e^{-n\,\mathrm{kl}\left(\frac{k}{n}\|p\right)} \tag{1.9}$$

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.

Example 12 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. We start with the following lemma.

Lemma 13. *Let $X_1, \ldots, X_n$ 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 1.9. □

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

Theorem 14 (kl inequality)**.** *Let $X_1, \ldots, X_n$ 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*

$$\mathbb{P}\left\{\mathrm{kl}(\hat{p}\|p) \geq \varepsilon\right\} \leq (n+1)e^{-n\varepsilon}. \tag{1.10}$$

Proof. By Markov's inequality and Lemma 13:

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

□

1.4.1 Comparison with Hoeffding's inequality

Is kl inequality tighter or loser than Hoeffding's inequality? In order to understand that we will use a couple of useful relaxations of the kl divergence.

Lemma 15 ((Cover and Thomas, 2006, Lemma 11.6.1))**.**

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

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

Corollary 16 (Pinsker's inequality)**.**

$$\mathrm{kl}(p\|q) \geq \frac{1}{2}\left(|p-q| + |(1-p)-(1-q)|\right)^2 = 2(p-q)^2. \tag{1.11}$$

Lemma 17. *If* $\mathrm{kl}(p\|q) \leq \varepsilon$ *then*

$$q \leq p + \sqrt{2p\varepsilon} + 2\varepsilon. \tag{1.12}$$

This lemma follows from the fact that for $q > p$ we have $\mathrm{kl}(p\|q) \geq (q-p)^2/(2q)$.

With these relaxations in hand we are ready to make the comparison. By denoting the right hand side of Hoeffding's inequality for binary variables (1.5) by δ, we obtain that Hoeffding's inequality assures that with probability greater than $1-\delta$:

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

By denoting the right hand side of kl inequality (1.10) by δ, we obtain that with probability greater than $1-\delta$:

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

By applying the first relaxation (1.11), we obtain that with probability greater than $1-\delta$:

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

Thus, the kl inequality is always at least as good as Hoeffding's inequality up to the $\ln(n+1)$ factor. However, if we use the tighter relaxation (1.12), 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 we used. Some improvements on the analysis presented in this section can be found in (Maurer, 2004, Langford, 2005).

Chapter 2

Supervised Learning

In this chapter we derive a number of generalization bounds for supervised learning. We start with a formal setting of supervised learning, formulate it in the language of probability theory, and then provide the analysis.

2.1 Supervised Learning Setting

We start with a bunch of notations.

- $\mathcal{X}$ - 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.

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. h is used to predict the unobserved Y and the quality of h is judged by the quality of predictions, as defined below.

We define:

- $\ell(Y', Y)$ - 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 (or error) of h on S. Sometimes, when it is clear from the context, we will omit S and write $\hat{L}(h)$.
- $L(h) = \mathbb{E}\left[\ell(h(X), Y)\right]$ - the expected loss (or error) of h, where the expectation is taken with respect to $p(X, Y)$.

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

2.2 Generalization Bound for a Single Hypothesis

We start with the simplest case, where we have a sample S and 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 $\hat{L}(h, S)$? Let $Z_i = \ell(h(X_i), Y_i)$ be the loss of h on the sample (X_i, Y_i). Note that Z_i is a random variable and that since (X_i, Y_i)-s are sampled i.i.d., $Z_1, \ldots, Z_n$ are i.i.d. random variables with $\mathbb{E}[Z_i] = L(h)$. And we also have $\hat{L}(h, S) = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Thus we have the following result.

Theorem 18. *Assume that ℓ is bounded in the $[0, 1]$ interval ($\ell(Y', Y) \in [0, 1]$ for all Y', Y), then for a single h and any $\delta \in (0, 1)$ we have:*

$$\mathbb{P}\left\{ L(h) > \hat{L}(h, S) + \sqrt{\frac{\ln \frac{1}{\delta}}{2n}} \right\} \leq \delta \tag{2.1}$$

and

$$\mathbb{P}\left\{ \left| L(h) - \hat{L}(h, S) \right| > \sqrt{\frac{\ln \frac{2}{\delta}}{2n}} \right\} \leq \delta. \tag{2.2}$$

Proof. For (2.1) take $\varepsilon = \sqrt{\frac{\ln \frac{1}{\delta}}{2n}}$ in (1.5) and rearrange the terms. Equation (2.2) follows in a similar way from the two-sided Hoeffding's inequality. Note that in (2.1) we have $\frac{1}{\delta}$ and in (2.2) we have $\frac{2}{\delta}$. □

There is an alternative way to read equation (2.1) - with probability greater than $1 - \delta$ we have:

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

In other words, with probability greater than $1-\delta$ the sample S is *representative* of the performance of h and $L(h)$ is close to $\hat{L}(h, S)$ up to $\sqrt{\frac{\ln \frac{2}{\delta}}{2n}}$ and with probability at most δ the sample S is not representative of the performance of h and we can say nothing about $L(h)$ by observing $\hat{L}(h, S)$.

Theorem 18 is directly analogous to the problem of estimating a bias of a coin based on coin flip outcomes. Lets say that we have a fair coin, but we do not know that this is a fair coin. With high probability, if we flip the coin 1000 times the empirical bias will be close to $\frac{1}{2}$, but there is always a small probability that we flip the coin 1000 times and get all heads or other non-representative outcome for this coin. 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 growth of the sample size n.

Whether we use the one-sided bound (2.1) or the two-sided bound (2.2) depends on the situation. Generally we care about the upper bound on the expected performance of the prediction rule (2.1), however, if we want to show that we did the best we could we may need to use (2.2).

2.3 Generalization Bound for Finite Hypothesis Classes

Taking a single prediction rule and estimating its expected error is a validation procedure. In learning we operate with multiple prediction rules and pick the best one based on the sample (like the best separating hyperplane in SVMs). So what can we say about the expected error $L(h)$, also called generalization error, when h was selected from a hypothesis class $\mathcal{H}$? The tricky point is that S may be representative for h_1, but not representative for h_2 and vice versa, and without any additional knowledge the best we can do is to apply a union bound.

Theorem 19. *Assume that ℓ is bounded in the $[0, 1]$ interval and that $|\mathcal{H}| = M$. Then for any $\delta \in (0, 1)$ we have:*

$$\mathbb{P}\left\{ \exists h \in \mathcal{H} : L(h) > \hat{L}(h, S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}} \right\} \leq \delta. \tag{2.3}$$

Proof.

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

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

Another way of reading Theorem 19 is - with probability greater than $1-\delta$ for all $h \in \mathcal{H}$:

$$L(h) \leq \hat{L}(h,S) + \sqrt{\frac{\ln \frac{M}{\delta}}{2n}}, \tag{2.4}$$

which means that no matter which h from $\mathcal{H}$ is returned by the algorithm, with high probability we have guarantee (2.4) on its expected performance. The price that we paid for considering M hypotheses instead of a single one is $\ln M$. Note that it grows only logarithmically with M!

Similar to theorem 18 it is possible to derive a two-sided bound on the error.

2.4 Occam's Razor Bound

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

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

where δ is the probability that things go wrong and $L(h)$ happens to be far away from $\hat{L}(h,S)$ because S is not representative for the performance of h. But there is an interdependence 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 will 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 go completely wrong. The idea behind Occam's Razor bound is to distribute this budget unevenly among the hypotheses in $\mathcal{H}$.

Theorem 20. *Let ℓ be bounded in $[0,1]$, let $\mathcal{H}$ be a countable hypothesis set and let $p(h)$ be such that $\sum_{h\in\mathcal{H}} p(h) \leq 1$ and $p(h)$ is independent of the sample. Then:*

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

Proof.

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

where the first inequality is the union bound, the second inequality is by Hoeffding's inequality, and the last inequality is by our assumption on $p(h)$. Note that $p(h)$ has to be selected before we observe the sample (or, in other words, independently of the sample), otherwise the second inequality does not hold. □

Another way of reading Theorem 20 is that with probability greater than $1-\delta$, for all $h \in \mathcal{H}$:

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

Again, note that the bound on $L(h)$ depends both on $\hat{L}(h,S)$ and on $p(h)$. 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 $p(h)$, where $p(h)$ can be interpreted as a complexity measure or a prior belief. Also, note that $p(h)$ can be designed arbitrarily, but it should be independent of the sample S. If $p(h)$ 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 $p(h)$ after observing S! Some considerations behind the choice of $p(h)$ are provided in Section 2.4.1.

2.4.1 Applications of Occam's Razor bound

We now consider two applications of Occam's Razor bound.

Generalization bound for finite hypotheses spaces

An immediate corollary from Occam's razor bound is the generalization bound for finite hypotheses classes that we have already seen earlier in the course.

Corollary 21. *Let $\mathcal{H}$ be a finite hypotheses class of size M, then*

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

Proof. We set $p(h) = \frac{1}{M}$ (which means that we distribute the confidence budget δ uniformly among the hypotheses in $\mathcal{H}$) and apply Theorem 20. □

Generalization bound for binary decision trees

Theorem 22. *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*

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

Proof. We first note that $|\mathcal{H}_d| = 2^{2^d}$. We define $p(h) = \frac{1}{2^{d(h)+1}} \frac{1}{2^{2^{d(h)}}}$. The first part of $p(h)$ distributes 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 $p(h)$ distributes confidence budget uniformly within $\mathcal{H}_d$. Since $\sum_{d=0}^{\infty} \frac{1}{2^{d+1}} = 1$, the assumption $\sum_{h\in\mathcal{H}} p(h) \leq 1$ is satisfied. The result follows by application of Theorem 20. □

Note that the bound depends on $\ln\left(\frac{1}{p(h)\delta}\right)$ and the dominating term in $\frac{1}{p(h)}$ 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 does not make 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.

2.5 Vapnik-Chervonenkis (VC) Generalization Bound

Please, see (Abu-Mostafa et al., 2012, Chapter 2 and Appendix).

2.6 VC Generalization Bound for SVMs

In this section we derive margin-based generalization bound for Support Vector Machines. The analysis is based on the following result that follows from (Abu-Mostafa et al., 2012, Theorem 2.5 and Problem 2.5).

Theorem 23 (VC generalization bound)**.** *Let $\mathcal{H}$ be a hypotheses class with VC-dimension $d_{VC}(\mathcal{H}) = d_{VC}$. Then:*

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

And we will use the following theorem without a proof.

Theorem 24 ((Abu-Mostafa et al., 2015, Theorem 8.5))**.** *Let $\mathcal{H}_\rho = \{(w,b) : \|w\| \geq 1/\rho\}$ be the space of hyperplanes described by w and b, where $\|w\| \geq 1/\rho$. Furthermore, assume that the input space $\mathcal{X}$ is a ball of radius R in $\mathbb{R}^d$ (so that $\|x\| \leq R$ for all $x \in \mathcal{X}$). Then:*

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

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

2.6.1 Analysis

For the analysis we make a simplifying assumption that $R = 1$. The analysis for general R is left as a home exercise.

Theorem 25. *Assume that $R = 1$. Let $\mathcal{H}$ be the space of linear separators $h = (w,b)$. Then*

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

Proof. We start by noting that Theorem 24 is interesting when $d_{VC}(\mathcal{H}_\rho) < d+1$ (because otherwise we can use the general bound for $\mathcal{H}$) and that the VC-dimension is an integer number. 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 $\mathcal{H}_i = \mathcal{H}_{\rho=i}$. By Theorem 24 we have $d_{VC}(\mathcal{H}_i) = i+1$ and then by Theorem 23:

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

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 $\mathcal{H}_i \setminus \mathcal{H}_{i-1}$ 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 $h = (w,b) \in \mathcal{H}_i \setminus \mathcal{H}_{i-1}$ then $\lceil\|w\|^2\rceil = i$. The remainder of

the proof follows the same lines as the proof of Occam's razor bound:

$$\mathbb{P}\left\{\exists h \in \mathcal{H} : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+\lceil\|w\|^2\rceil}+1\right)(1+\lceil\|w\|^2\rceil)\lceil\|w\|^2\rceil/\delta\right)}{n}}\right\}$$

$$= \mathbb{P}\left\{\exists h \in \bigcup_{i=1}^{d} \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(\left((2n)^{1+\lceil\|w\|^2\rceil}+1\right)(1+\lceil\|w\|^2\rceil)\lceil\|w\|^2\rceil/\delta\right)}{n}}\right\}$$

$$= \sum_{i=1}^{d} \mathbb{P}\left\{\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+\lceil\|w\|^2\rceil}+1\right)(1+\lceil\|w\|^2\rceil)\lceil\|w\|^2\rceil/\delta\right)}{n}}\right\}$$

$$= \sum_{i=1}^{d} \mathbb{P}\left\{\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+i}+1\right)(1+i)\,i/\delta\right)}{n}}\right\}$$

$$= \sum_{i=1}^{d} \mathbb{P}\left\{\exists h \in \mathcal{H}_i \setminus \mathcal{H}_{i-1} : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+i}+1\right)/\delta_i\right)}{n}}\right\}$$

$$\leq \sum_{i=1}^{d} \mathbb{P}\left\{\exists h \in \mathcal{H}_i : L(h) > \hat{L}(h,S) + \sqrt{\frac{8\ln\left(2\left((2n)^{1+i}+1\right)/\delta_i\right)}{n}}\right\}$$

$$\leq \sum_{i=1}^{d} \delta_i = \sum_{i=1}^{d} \frac{1}{i(i+1)}\delta = \delta\sum_{i=1}^{d} \frac{1}{i(i+1)} \leq \delta\sum_{i=1}^{\infty} \frac{1}{i(i+1)} = \delta.$$

□

2.7 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 26. *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*

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

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 $h_i \in \mathcal{F}_k(S)$ the empirical error $\hat{L}(h^*(f_i,S),S) = 0$.

Let $p(k)$ and $p(i)$ be uniform. Then:

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

□

Corollary 27. *Under the assumptions of Theorem 26, 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 $\mathbb{P}\left\{\sup_h L(h) - \hat{L}(h,S) \geq 0.125\right\} < 0.125$. Then

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

which is in contradiction with Theorem 26. □

2.8 PAC-Bayesian Analysis

PAC-Bayesian analysis is an alternative way to derive generalization bounds for infinite hypothesis classes. PAC-Bayesian bounds are defined for *randomized classifiers*, which we define below. PAC-Bayesian generalization bounds are based on *change of measure* inequality, which acts as a replacement for the union bound when we are working with infinite sets. (In the home assignment you will verify that change of measure inequality is slightly tighter than the union bound.)

Definition 28 (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 denoted by $L(\rho)$, the empirical loss of ρ is denoted by $\hat{L}(\rho,S)$ and they are defined by:

$$L(\rho) = \mathbb{E}_{(x,y)\sim p, h\sim\rho}\left[\ell(y,h(x))\right] = \mathbb{E}_{h\sim\rho}[L(h)] = \langle L,\rho\rangle = \begin{cases} \sum_{h\in\mathcal{H}} L(h)\rho(h), & \textit{Discrete } \mathcal{H} \\ \int_{\mathcal{H}} L(h)\rho(h)dh, & \textit{Continuous } \mathcal{H} \end{cases}.$$

$$\hat{L}(\rho,S) = \mathbb{E}_{h\sim\rho}[\hat{L}(h,S)] = \langle \hat{L},\rho\rangle$$

There is a large number of different PAC-Bayesian inequalities. We start with the classical one.

Theorem 29 (PAC-Bayes-kl inequality)**.** *For any "prior" distribution π over $\mathcal{H}$, for all randomized classifiers (distributions) ρ simultaneously:*

$$\mathbb{P}\left\{\mathrm{kl}(\hat{L}(\rho,S)\|L(\rho)) \geq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}\right\} \leq \delta.$$

By using Pinsker's inequality 16 we can write this result in a more digestible (although weaker) form - with probability greater than $1-\delta$ for all ρ simultaneously

$$L(\rho) \leq \hat{L}(\rho, S) + \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln \frac{n+1}{\delta}}{2n}}.$$

Note that when $\rho = \pi$ the KL term is zero and we recover generalization bound for a single hypothesis. (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.) Also note that if $\mathcal{H}$ is finite and π is uniform then $\mathrm{KL}(\rho\|\pi) = \ln|\mathcal{H}| - \mathrm{H}(\rho) \leq \ln|\mathcal{H}|$ and we recover generalization bound for finite hypothesis sets with a slight improvement by $-\mathrm{H}(\rho)$.

To get more intuition we decompose the KL-divergence:

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

Thus, PAC-Bayesian inequality encourages us to pick 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 $\pi(h) = \pi(h')$)

PAC-Bayesian analysis has the following relation and differences with Bayesian learning and with VC analysis / Radamacher complexities, which are complimentary to each other.

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 many cases

In a sence, 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.

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

Theorem 30 (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. Note that there is nothing probabilistic in the statement of the theorem - it is a deterministic result. □

Different PAC-Bayesian inequalities are obtained by different choices of the function $f(h)$. In particular, for PAC-Bayes-kl inequality we pick $f(h) = n\,\mathrm{kl}(\hat{L}(h, S)\|L(h))$. The key consideration in the choice of $f(h)$ is the ability to bound $\mathbb{E}\left[e^{f(h)}\right]$. We have already done it for the kl divergence in Lemma 13. Note that we have also done it for $f(h) = \lambda\left(L(h) - \hat{L}(h, S)\right)$ in the Hoeffding's lemma 5, so this could be another possible choice that would lead to PAC-Bayes-Hoeffding inequality. And there are other possible choices (Seldin et al., 2012).

Proof of Theorem 29. We note that $\mathrm{kl}(p\|q)$ is convex in the pair p, q (Cover and Thomas, 2006, Theorem 2.7.2) and, therefore

$$\begin{aligned}
n\,\mathrm{kl}(\hat{L}(\rho, S)\|L(\rho)) &= n\,\mathrm{kl}\left(\mathbb{E}_\rho\left[\hat{L}(h, S)\right]\middle\|\mathbb{E}_\rho\left[L(h)\right]\right) \\
&\leq \mathbb{E}_\rho\left[n\,\mathrm{kl}(\hat{L}(h, S)\|L(h))\right] \\
&\leq \mathrm{KL}(\rho\|\pi) + \ln \mathbb{E}_\pi\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right],
\end{aligned}$$

where the last step is by change of measure inequality. Note that this statement is deterministic and there is no randomness involved. By normalizing by n and taking $\mathrm{KL}(\rho\|\pi)$ to the other side we obtain:

$$\begin{aligned}
\mathbb{P}\left\{\mathrm{kl}(\hat{L}(\rho, S)\|L(\rho)) \geq \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}\right\} &= \mathbb{P}\left\{n\,\mathrm{kl}(\hat{L}(\rho, S)\|L(\rho)) - \mathrm{KL}(\rho\|\pi) \geq \ln\frac{n+1}{\delta}\right\} \\
&\leq \mathbb{P}\left\{\ln \mathbb{E}_\pi\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right] \geq \ln\frac{n+1}{\delta}\right\} \\
&= \mathbb{P}\left\{\mathbb{E}_\pi\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right] \geq \frac{n+1}{\delta}\right\} \\
&\leq \delta \times \frac{\mathbb{E}_S\left[\mathbb{E}_\pi\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right]\right]}{n+1} \qquad (2.5) \\
&= \delta \times \frac{\mathbb{E}_\pi\left[\mathbb{E}_S\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right]\right]}{n+1} \qquad (2.6) \\
&\leq \delta \times \frac{\mathbb{E}_\pi\left[n+1\right]}{n+1} \qquad (2.7) \\
&= \delta.
\end{aligned}$$

The key steps are:

- Step (2.5) is by Markov's inequality.
- In equation (2.6) we can exchange the expectations since π *is independent of* S. This is the step where the independence is crucial.
- Step (2.7) is by Lemma 13. This is the step, where for a different choice of f we could use another bound on expectation of the moment generating function, for example, Hoeffding's Lemma 5.

And, we emphasize once again, that the most crucial point was that the change of measure inequality relates all posterior distributions ρ to a single prior distribution π in a deterministic manner. Thus, there is only one random quantity that we had to bound which is $\mathbb{E}_\pi\left[e^{n\,\mathrm{kl}(\hat{L}(h,S)\|L(h))}\right]$. □

2.8.1 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).

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.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory.* Wiley-Interscience, 2nd edition, 2006.

John Langford. Tutorial on practical prediction theory for classification. *Journal of Machine Learning Research*, 2005.

Andreas Maurer. A note on the PAC-Bayesian theorem. www.arxiv.org, 2004.

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.