A survey on concentration inequalities

Martín Prado

November 2023

Universidad de los Andes – Bogotá Colombia

Contents

1 Introduction

1.1 Basic inequalities and theorems

Theorem 1.1 (Markov's inequality)**.** For a random variable X with $\mathbf{P}\{X < 0\} = 0$ and $t > 0$, we have

$$\mathbf{P}\{X \geq t\} \leq \frac{\mathbf{E}\, X}{t}.$$

Proof. In the first place, note that

$$\begin{aligned} X &= X \cdot \mathbb{1}_{\{X \geq t\}} + X \cdot \mathbb{1}_{\{X < t\}} \\ &\geq \ t \cdot \mathbb{1}_{\{X \geq t\}} + 0, \end{aligned}$$

and thus,

$$\mathbf{E}\, X \geq t \cdot \mathbf{E}\, \mathbb{1}_{\{X \geq t\}} = t \cdot \mathbf{P}\{X \geq t\}.$$

$\square$

Theorem 1.2 (Chebyshev's inequality)**.** For $t > 0$, a random variable X with mean $\mu = \mathbf{E}\, X$ and variance $\sigma^2 = \mathbf{Var}\, X$, we have

$$\mathbf{P}\{|X - \mu| \geq t\} \leq \frac{\sigma^2}{t^2}.$$

Proof. We apply Markov's inequality to the non-negative random variable $Y = |X - \mu|^2$ in order to obtain the desired result

$$\mathbf{P}\{|X - \mu| \geq t\} = \mathbf{P}\{|X - \mu|^2 \geq t^2\} \leq \frac{\mathbf{E}\left[(X - \mu)^2\right]}{t^2} = \frac{\sigma^2}{t^2}.$$

$\square$

1.2 Why bother?

The concentration inequalities are used to obtain information on how a random variable is distributed at some specific places of its domain. In the most common scenarios, these inequalities will be used to quantify how concentrated a random variable around its center. In other words, how fast the probability decays as we move towards the tails. For example,

$$\mathbf{P}\{|X-\mu| \geq t\} < f(t) << 1.$$

A concentration inequality is specially useful when this probability cannot be calculated at a low computational cost or estimated with high precision. The following will illustrate a case where using concentration inequalities achieves the best results.

1.2.1 Coin Tossings

A coin tossing game is fair if the chances of winning are equal to the chances of losing. We can verify from a sample of N games that the game is not rigged if the number of heads in the sample is not very distant from the average $N/2$. However, there's a chance that one may classify the coin as rigged, even when the coin is fair. By the *Law of Large Numbers*, we know that the larger the sample, the less likely it is to obtain a false positive. But let's ask ourselves how fast this probability converges to 0.

Let $S_N \sim \mathrm{Bi}(N, 1/2)$ denote the number of heads in a fair coin tossing game. Then,

$$\mu = \mathbf{E}\, S_N = \frac{N}{2}, \qquad \sigma^2 = \mathbf{Var}\, S_N = \frac{N}{4}.$$

For a fixed $\varepsilon > 0$, we may classify a coin tossing game as rigged if, after N trials, the ratio of success falls outside the interval $[0, \frac{1+\varepsilon}{2}]$

$$S_N \geq \mu + \frac{\varepsilon}{2}N = \frac{1+\varepsilon}{2}N.$$

It's clear that calculating the exact probability of the previous event for any N, ε is a very demanding task computationally. The Chebyshev's inequality 1.2 gives us a "good-enough" result for this problem,

$$\mathbf{P}\left\{S_N \geq \mu + \frac{\varepsilon}{2}N\right\} \leq \mathbf{P}\left\{|S_N - \mu| \geq \frac{\varepsilon}{2}N\right\} \leq \sigma^2 \frac{4}{\varepsilon^2 N^2} = \frac{1}{\varepsilon^2 N}.$$

Therefore, the probability of bad events tends to 0 at least linearly with the number of games.

1.2.2 Central Limit Theorem

The proof of the following three theorems can be found in Boucheron et al. (2003)

Theorem 1.3. Let X_i be a i.i.d. sample. Let $S_N = \sum_{i=1}^{N} X_i$, with mean $\mu = \mathbf{E}\, S_N$ and variance $\sigma^2 = \mathbf{Var}\, S_N$. If

$$Z_N = \frac{S_N - N \cdot \mathbf{E}\, X_i}{\sqrt{N \cdot \mathbf{Var}\, X_i}} = \frac{S_N - \mu}{\sqrt{N}\sigma},$$

then,

$$Z_N \to Z \sim \mathcal{N}(0,1), \text{ in distribution.}$$

□

Theorem 1.4 (Tails of the Normal Distribution)**.** Let $Z \sim \mathcal{N}(0,1)$, for $t > 0$ we have

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right) \le \mathbf{P}\{Z \ge t\} \le \frac{1}{t} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-t^2}{2}\right).$$

□

With that in mind, we might naively assume that better bounds can be obtained by using the previous theorem. For a large enough N we can say that for the coin tossing,

$$Z_N = \frac{S_N - N/2}{\sqrt{N/4}}$$

$$\implies \mathbf{P}\left\{S_N \ge \frac{1+\varepsilon}{2} N\right\} = \mathbf{P}\left\{Z_N \ge \varepsilon\sqrt{N}\right\} \sim \mathbf{P}\left\{Z \ge \varepsilon\sqrt{N}\right\}.$$

However, this raises the question of whether we can draw the following conclusion from Theorem 1.4:

$$\mathbf{P}\left\{S_N \ge \frac{1+\varepsilon}{2} N\right\} \le \frac{1}{\varepsilon\sqrt{N}} \frac{1}{\sqrt{2\pi}} \exp\left(\frac{-\varepsilon^2 \cdot N}{2}\right).$$

Unfortunately, the answer is no. The following theorem will show why.

Theorem 1.5 (Convergence Rate for Central Limit Theorem)**.** For Z_N, Z in Theorem 1.3, we have:

$$|\mathbf{P}\{Z_N \ge t\} - \mathbf{P}\{Z \ge t\}| = O(\tfrac{1}{\sqrt{N}}).$$

□

Since the approximation error of the Central Limit Theorem is of greater order than the normal bounds, the previous results cannot be taken into account.

In the context of coin tossing, this may not matter at all because the linear bound obtained using Chebyshev's inequality indicates that the probability of wrongly classifying a fair coin as a rigged coin converges at least linearly to zero. Even the Central Limit Theorem shows, in a less precise way, this convergence. However, for some specific problems in statistics, these basic tools are not precise enough to solve them. The main objective of this project is to study different ideas that improve these bounds and show examples where they can be used.

1.3 Cantelli's inequality

We can start with a small modification of the Chebyshev's bound for the one-sided tails

Theorem 1.6 (Cantelli's Inequality)**.** For $t > 0$, a random variable X with mean $\mu = \mathbf{E}\, X$ and variance $\sigma^2 = \mathbf{Var}\, X$, we have

$$\mathbf{P}\{X - \mu \geq t\} \leq \frac{\sigma^2}{t^2 + \sigma^2}.$$

Proof. In the first place note that,

$$\mathbf{P}\{Y \geq s\} \leq \mathbf{P}\{Y \geq s\} + \mathbf{P}\{Y \leq s\} = \mathbf{P}\{|Y| \geq s\} = \mathbf{P}\{Y^2 \geq s^2\}. \qquad (\star)$$

Let $u \geq 0$, define $Y = X - \mu + u$ and $s = t + u$ to obtain

$$\mathbf{P}\{X - \mu \geq t\} = \mathbf{P}\{X - \mu + u \geq t + u\} = \mathbf{P}\{Y \geq s\}.$$

We use $(\star)$ and Markov's inequality (1.1) on Y^2 to conclude,

$$\mathbf{P}\{Y \geq s\} \overset{(\star)}{\leq} \mathbf{P}\{Y^2 \geq s^2\} \overset{(1.1)}{\leq} \frac{\mathbf{E}\left[(X - \mu + u)^2\right]}{(t+u)^2}.$$

By linearity of expectation,

$$\mathbf{E}\left[(X - \mu + u)^2\right] = \mathbf{E}\left[(X - \mu)^2\right] + 2u \cdot \underbrace{\mathbf{E}\,(X - \mu)}_{0} + E(u^2) = \sigma^2 + u^2.$$

Finally, we choose an optimal $u = \frac{\sigma^2}{t}$ to conclude

$$\mathbf{P}\{X - \mu \geq t\} \leq \frac{\sigma^2 + u^2}{(t+u)^2} = \frac{\sigma^2 + \sigma^4/t^2}{(t + \sigma^2/t)^2} = \frac{\sigma^2\left(\frac{t^2+\sigma^2}{t^2}\right)}{\left(\frac{t^2+\sigma^2}{t}\right)^2} = \frac{\sigma^2}{t^2 + \sigma^2}.$$

□

On the other hand, the two-sided tail inequality, Cantelli's inequality is not always better than Chebyshev,

Corollary 1.6.1 (Two-sided Cantelli inequality)**.**

$$\mathbf{P}\{|X - \mu| \geq t\} \leq \frac{2\sigma^2}{t^2 + \sigma^2}.$$

In fact, this bound is only better than Chebyshev's $t^2 + \sigma^2 \geq 2t^2$, or equivalently, when $\sigma^2 \geq t^2$. However, in this case both formulas provide bounds greater than 1, and thus, are useless. Therefore, the conclusion is that, in general, Chebyshev's inequality is better for two-sided tails and Cantelli is for one-sided tails.

2 Exponential Inequalities

Even if we are satisfied with the linear convergence rate provided by Chebyshev's inequality or the improvement of one sided tails given by Cantelli's inequality, there is a simple but powerful modification we can make to Markov's inequality that will greatly improve both bounds. The following result will provide the main idea from which most of the exponential inequalities are derived.

Theorem 2.1 (MGF inequality)**.** Let X_i be a finite sequence of independent random variables and let $S_N := \sum_{i=1}^N a_i X_i$. Let $\lambda > 0$. The following inequality holds,

$$\mathbf{P}\left\{S_N \geq t\right\} \leq e^{-\lambda t} \cdot \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda a_i X_i}.$$

Proof. Let $\lambda > 0$, using Markov's inequality (Theorem 1.1) we assert that since $x \mapsto e^{\lambda x}$ is a non-decreasing function,

$$\mathbf{P}\left\{S_N \geq t\right\} = \mathbf{P}\left\{e^{\lambda S_N} \geq e^{\lambda t}\right\} \leq e^{-\lambda t} \cdot \mathbf{E}\, \exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right).$$

Since X_i are independent, the MGF of S_N is the product of MGFs of each X_i:

$$\mathbf{E}\, \exp\left(\lambda \sum_{i=1}^{N} a_i X_i\right) = \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda a_i X_i}$$

$$\implies \mathbf{P}\left\{S_N \geq t\right\} \leq e^{-\lambda t} \cdot \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda a_i X_i}.$$

$\square$

The following two theorems are examples on how we can obtain even tighter bounds than the ones we've already studied. In particular, these theorems can be obtained from the previous theorem and are considered, by some authors, as corollaries of the previous result.

Theorem 2.2 (Chernoff's inequality)**.** Let $X_i \sim \mathrm{Be}(p_i)$ be independent random variables. Define $S_N = \sum_{i=1}^N X_i$ and let $\mu = \mathbf{E}\, S_N$. Then, for $t > \mu$, we have

$$\mathbf{P}\left\{S_N \geq t\right\} \leq \left(\frac{\mu}{t}\right)^t e^{-\mu + t}.$$

Proof. In the first place, use Theorem 2.1 to assert that for a $\lambda > 0$ that

$$\mathbf{P}\{S_N \geq t\} \leq e^{-\lambda t} \cdot \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda X_i}.$$

Now it is left to bound every X_i individually. Using the inequality $1 + x \leq e^x$ we obtain

$$\mathbf{E}\, e^{\lambda X_i} = e^{\lambda} p_i + (1 - p_i) = 1 + (e^{\lambda} - 1)p_i \leq \exp(e^{\lambda} - 1)e^{p_i}.$$

Finally, we plug this inequality on the equation to conclude that

$$e^{-\lambda t} \cdot \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda X_i} \leq e^{-\lambda t} \cdot \prod_{i=1}^{N} \exp((e^{\lambda} - 1)p_i) = e^{-\lambda t} \exp((e^{\lambda} - 1)\mu).$$

By using the substitution $\lambda = \ln(t/\mu)$ we obtain the desired result,

$$\mathbf{P}\{S_N \geq t\} \leq \left(\frac{\mu}{t}\right)^t \exp\left(\frac{\mu t}{\mu} - \mu\right) = \left(\frac{\mu}{t}\right)^t e^{-\mu + t}.$$

□

Another exponential inequality that is derived using a similar technique is Hoeffding's inequality:

Theorem 2.3 (Hoeffding's inequality)**.** Let $X_1, \ldots, X_N$ be independent random variables, such that $X_i \in [a_i, b_i]$ for every $i = 1, \ldots, N$. Define $S_N = \sum_{i=1}^{N} X_i$ and let $\mu = \mathbf{E}\, S_N$. Then, for every $t > 0$, we have

$$\mathbf{P}\{S_N \geq \mu + t\} \leq \exp\left(\frac{-2t^2}{\sum (a_i - b_i)^2}\right).$$

Proof. Since $x \mapsto e^x$ is a convex function, it follows that, for a random variable $X \in [a, b]$:

$$e^{\lambda X} \leq \frac{e^{\lambda a}(b - X)}{b - a} + \frac{e^{\lambda b}(X - a)}{b - a}, \quad a \leq b.$$

Next, take expectations on both sides of the equation to obtain:

$$\mathbf{E}\, e^{tX} \leq \frac{(b - \mathbf{E}\, X) \cdot e^{\lambda a}}{b - a} + \frac{(\mathbf{E}\, X - a) \cdot e^{\lambda b}}{b - a}.$$

To simplify the expression, let $\alpha = (\mathbf{E}\, X - a)/(b - a)$, $\beta = (b - \mathbf{E}\, X)/(b - a)$ and $u = \lambda(b - a)$. Since $a < \mathbf{E}\, X < b$, it follows that α and β are positive. Also, note that,

$$\alpha + \beta = \frac{\mathbf{E}\, X - a}{b - a} + \frac{b - \mathbf{E}\, X}{b - a} = \frac{b - a}{b - a} = 1.$$

Now,

$$\ln \mathbf{E}\, e^{\lambda X} \leq \ln(\beta e^{-\alpha u} + \alpha e^{\beta u}) = -\alpha u + \ln(\beta + \alpha e^u).$$

This function is differentiable with respect to u.

$$\begin{aligned} L(u) &= -\alpha u + \ln(\beta + \alpha e^{u}), \\ L'(u) &= -\alpha + \frac{\alpha}{\alpha + \beta e^{-u}}, \\ L''(u) &= \frac{\alpha}{\alpha + \beta e^{-u}} \cdot \frac{\beta e^{-u}}{\alpha + \beta e^{-u}}. \end{aligned}$$

Note that if $x = \frac{\alpha}{\alpha+\beta e^{-u}} \leq 1$, then $L''(u) = x(1-x) \leq \frac{1}{4}$. Remember that $\alpha + \beta = 1$. Now, by expanding the Taylor series we obtain,

$$\begin{aligned} L(u) &= L(0) + uL'(0) + \frac{1}{2}u^2 L''(u) \\ &= \ln(\beta + \alpha) + u\left(-\alpha + \frac{\alpha}{\alpha+\beta}\right) + \frac{1}{2}u^2 L''(u) \\ &= \frac{1}{2}u^2 L''(u) \\ &\leq \frac{1}{8}\lambda^2 (b-a)^2. \end{aligned} \qquad (\star)$$

Finally, use the inequality from Theorem 2.1 to conclude that

$$\begin{aligned} \mathbf{P}\{S_N - \mu \geq t\} \;\; &\leq \;\; e^{-\lambda t} \prod_{i=1}^{N} \mathbf{E}\, e^{\lambda X_i} \\ &\leq^{(\star)} \; e^{-\lambda t} \exp\left(\frac{1}{8} t^2 \sum_{i=1}^{N} (b_i - a_i)^2\right). \end{aligned}$$

□

Corollary 2.3.1. Let $X_1, \ldots, X_N$ be independent random Bernoulli variables such that $X_i \sim \mathrm{Be}(p_i)$, then

$$\mathbf{P}\left\{\sum_{i=1}^{N}(X_i - p_i) \geq t\right\} \leq \exp\left(\frac{-2t^2}{N}\right).$$

.

□

Returning to the coin tossing problem, we can now make a stronger assertion of the rate of convergence of a false negative classification using Hoeffding inequality:

$$\mathbf{P}\left\{S_N - \frac{N}{2} \geq \frac{\varepsilon}{2} N\right\} \leq \exp\left(-\varepsilon^2 N\right).$$

2.1 Uniform Law of Large Numbers

For any probability measure P on the real line and $t \in \mathbb{R}$, define P_n as the empirical probability measure obtained from an independent sample $X_1, \dots, X_n$ of P, that is:

$$P_n(t) = P_n(-\infty, t) = n^{-1} \cdot \sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}}.$$

From the law of large numbers we know that for a fixed t, $P_n(t)$ converges to $P(t)$ with probability 1. However we can formulate a stronger statement on this convergence. The first application of concentration inequalities we are going to explore is the uniform law of large numbers, which states the following:

Theorem 2.4 (Glivenko-Cantelli Theorem)**.** For P, P_n and $t \in \mathbb{R}$,

$$\|P_n - P\| = \sup_{t \in \mathbb{Q}} |P_n(t) - P(t)| \xrightarrow{p} 0.$$

Proof. The proof, adapted from Pollard (1984), consists of 5 steps. At first instance, the author clarifies that we can impose the condition $t \in \mathbb{Q}$ to avoid problems with measurability. The author later proves that the theorem is true t is allowed to vary in $\mathbb{R}$, but for practical purposes, we will only prove it for rationals. Another remark the author makes is that this result from the real line can be later generalized for some classes of "polynomial discrimination", and we will cover more about this in the final section.

First Symmetrization

In the first place, define P'_n as the empirical measure obtained from an independent but identical sample $X'_1, \dots, X'_n$ of P. Note that for any fixed t, $P_n(t)$ and $P'_n(t)$ are random variables derived from their respective samples which satisfy that

$$\mathbf{E}\, P_n(t) = \mathbf{E}\, P'_n(t) = P(t).$$

We will bound the concentration of $\|P_n - P'_n\|$ first, which will later result in a bound for $\|P_n - P\|$ at the end of the following lemma.

For now, fix a value for $\varepsilon > 0$, and keep in mind that $Z = P_n - P$, $Z' = P'_n - P$, $\alpha = \frac{1}{2}\varepsilon$ and $\beta = \frac{1}{2}$. Also, for this case define $\mathscr{A} = \{(-\infty, t) : t \in \mathbb{R}\}$

Lemma 2.5. Let $\{Z(A)\}_{A \in \mathscr{A}}$ and $\{Z'(A)\}_{A \in \mathscr{A}}$ be independent and identical functions defined under the same collection of sets $\mathscr{A}$. Also, assume that there exist $\alpha, \beta > 0$ such that

$$\mathbf{P}\{|Z(A)| \le \alpha\} \ge \beta, \quad \forall A \in \mathscr{A}$$

It follows that, for any $\varepsilon > 0$,

$$\mathbf{P}\left\{\sup_{A \in \mathscr{A}} |Z(A)| > \varepsilon\right\} \le \beta^{-1} \mathbf{P}\left\{\sup_{A \in \mathscr{A}} |Z(A) - Z'(A)| > \varepsilon - \alpha\right\}.$$

Proof. Since Z, Z' are independent, it follows from the hypothesis that for any index $B \in \mathscr{A}$,

$$\mathbf{P}\{|Z'(B)| \leq \alpha | Z\} = \mathbf{P}\{|Z'(B)| \leq \alpha\} \geq \mathbf{P}\left\{\sup_{A \in \mathscr{A}} |Z'(A)| \leq \alpha\right\} \geq \beta.$$

Fix an index B in the set $\{A \in \mathscr{A} \ : \ |Z(A)| > \varepsilon\}$. Since

$$\mathbf{P}\{\sup_{A \in \mathscr{A}} |Z(A)| > \varepsilon\} \leq \mathbf{P}\{|Z(B)| > \varepsilon\},$$

it follows that,

$$\begin{aligned} \beta \cdot \mathbf{P}\left\{\sup_{A \in \mathscr{A}} |Z(A)| > \varepsilon\right\} &\leq \mathbf{P}\{|Z'(B)| \leq \alpha\} \cdot \mathbf{P}\{|Z(B)| > \varepsilon\} \\ \text{\small (Z, Z' are independent)} &= \mathbf{P}\{|Z'(B)| \leq \alpha \text{ and } |Z(B)| > \varepsilon\} \\ &\leq \mathbf{P}\{|Z(B) - Z'(B)| > \varepsilon - \alpha\} \\ &\leq \mathbf{P}\left\{\sup_{A \in \mathscr{A}} |Z(A) - Z'(A)| > \varepsilon - \alpha\right\}. \end{aligned}$$

□

Using Chevyshev's inequality (1.2) we know that the hypothesis is satisfied for the values of α and β we chose:

$$\forall t \in \mathbb{R} : \mathbf{P}\left\{|Z'(t)| \leq \alpha\right\} = \mathbf{P}\{|P_n(t) - P(t)| \leq \varepsilon\} \geq \frac{1}{2} = \beta, \quad \text{if } n \geq 8\varepsilon^{-2}.$$

Therefore, using the previous lemma, we conclude that

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \leq 2\mathbf{P}\{\|P_n - P_n'\| > \tfrac{1}{2}\varepsilon\}, \quad \text{if } n \geq 8\varepsilon^{-2}. \tag{2.1.1}$$

Second Symmetrization

The following trick will allow us to stop considering all of the $2n$ data points from the previous symmetrization, and will help us to create a simpler random variable. We will initially prove the trick for unidimensional random variables, but in chapter 4, we will generalize this proof for any kind on set on $\mathbb{R}^n$.

Lemma 2.6. Let $\sigma_1, \dots, \sigma_n$ be Rademacher random variables, that is $\mathbf{P}\{\sigma_i = 1\} = \mathbf{P}\{\sigma_i = -1\} = 1/2$. Let $Y_i = \mathbb{1}_{\{X_i' \in A\}} - \mathbb{1}_{\{X_i \in A\}}$, and note that,

$$\mathbf{P}\{Y_i = x\} = \mathbf{P}\{\sigma_i Y_i = x\}, \quad x \in \{-1, 0, 1\}.$$

Proof. In the first place, since X_i and X_i' are two independent and identical copies of the same distribution, the following equality holds:

$$\begin{aligned}\mathbf{P}\{Y_i = 1\} &= \mathbf{P}\{X_i \in A\}\mathbf{P}\{X_i' \notin A\}\\ &= \mathbf{P}\{X_i' \in A\}\mathbf{P}\{X_i \notin A\}\\ &= \mathbf{P}\{Y_i = -1\}.\end{aligned}$$

On the other hand, since σ_i is also independent of Y_i, it follows that

$$\begin{aligned}\mathbf{P}\{\sigma_i Y_i = 1\} &= \mathbf{P}\{Y_i = 1, \sigma_i = 1\} + \mathbf{P}\{Y_i = -1, \sigma_i = -1\}\\ &= \mathbf{P}\{Y_i = 1\}\mathbf{P}\{\sigma_i = 1\} + \mathbf{P}\{Y_i = -1\}\mathbf{P}\{\sigma_i = 1\}\\ &= \frac{1}{2}\mathbf{P}\{Y_i = 1\} + \frac{1}{2}\mathbf{P}\{Y_i = 1\}\\ &= \mathbf{P}\{Y_i = 1\} = \mathbf{P}\{Y_i = -1\} = \mathbf{P}\{\sigma_i Y_i = -1\}.\end{aligned}$$

Thus,

$$\mathbf{P}\{\sigma_i Y_i = \pm 1\} = \mathbf{P}\{Y_i = \pm 1\}, \quad \mathbf{P}\{\sigma_i Y_i = 0\} = \mathbf{P}\{Y_i = 0\}.$$

$\square$

It follows that since $P_n - P_n' = n^{-1}\sum_{i \le n} Y_i$,

$$\begin{aligned}\mathbf{P}\{\|P_n - P_n'\| > \tfrac{1}{2}\varepsilon\} &= \mathbf{P}\left\{\sup_{t\in\mathbb{Q}}\left|n^{-1}\sum_{i=1}^{n}\sigma_i Y_i\right| > \tfrac{1}{2}\varepsilon\right\}\\ &\le \mathbf{P}\left\{\sup_{t\in\mathbb{Q}}\left|n^{-1}\sum_{i=1}^{n}\sigma_i \mathbb{1}_{\{X_i<t\}}\right| > \tfrac{1}{4}\varepsilon\right\} \qquad (2.1.2)\\ &\quad + \mathbf{P}\left\{\sup_{t\in\mathbb{Q}}\left|n^{-1}\sum_{i=1}^{n}\sigma_i \mathbb{1}_{\{X_i'<t\}}\right| > \tfrac{1}{4}\varepsilon\right\}\\ &= 2\mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\}.\end{aligned}$$

where $P_n^\circ = n^{-1}\sum_{i\le n}\sigma_i \mathbb{1}_{\{X_i<t\}}$. Then, from equations 2.1.1, 2.1.2 we conclude that for $n \ge 8\varepsilon^{-2}$,

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \le 4\mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\}.$$

Maximal Inequality

$$-\infty \xleftarrow{\;t_0\;} X_{(1)} \xrightarrow{\;t_1\;} X_{(2)} \xrightarrow{\;t_2\;} X_{(3)} \xrightarrow{\;t_3\;} \cdots \xrightarrow{\;t_{n-1}\;} X_{(n)} \xrightarrow{\;t_n\;} \infty$$

For any given sample $X = X_1, \dots, X_n$, define $X_{(j)}$ as the j-th observation when we order the observations, and fix $t_j \in (X_{(j)}, X_{(j+1)}]$ for every $j \leq n$ as the picture above shows. Note that if $t \in (X_{(j)}, X_{(j+1)}]$, then $P_n^\circ(t) = P_n^\circ(t_j)$ because:

$$\begin{aligned}
P_n^\circ(t) &= n^{-1}\sum_{i=1}^{n} \sigma_i \mathbb{1}_{\{X_i<t\}}, \qquad\qquad t \in (X_{(j)}, X_{(j+1)}] \\
&= n^{-1}\sum_{i=j+1}^{n} \sigma_i \mathbb{1}_{\{X_{(i)}<t\}} + n^{-1}\sum_{i=1}^{j} \sigma_i \mathbb{1}_{\{X_{(i)}<t\}} \\
&= n^{-1}\sum_{i=j+1}^{n} \sigma_i \cdot 1 \qquad + \qquad 0 \\
&= P_n^\circ(t_j).
\end{aligned}$$

It follows that for some k, $\|P_n^\circ\| = |P_n^\circ(t_k)|$, and thus,

$$\begin{aligned}
\mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid X\} &\leq \sum_{j=0}^{n} \max_j \mathbf{P}\{|P_n^\circ(t_j)| > \tfrac{1}{4}\varepsilon \mid X\} \\
&\leq (n+1) \cdot \mathbf{P}\{|P_n^\circ(t_k)| > \tfrac{1}{4}\varepsilon \mid X\}.
\end{aligned} \tag{2.1.3}$$

Exponential Bounds

Since for any given sample, $\sigma \mathbb{1}_{X_i<t} \in [-1, 1]$, we can use Hoeffding's Inequality 2.3 to obtain the following inequality

$$\mathbf{P}\{|P_n^\circ(A)| > \tfrac{1}{4}\varepsilon\} \leq 2\exp\left(\frac{-2(n\varepsilon/4)^2}{4n}\right) = 2e^{-n\varepsilon^2/32}, \quad \forall A \in \mathscr{A}.$$

We use equation 2.1.3 to conclude

$$\mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid X\} \leq 2(n+1)e^{-n\varepsilon^2/32}. \tag{2.1.4}$$

Integration

Finally, applying the formula $P\{A\} = \mathbf{E}_X[\mathbf{P}\{A|X\}]$, we conclude that

$$\begin{aligned}
\mathbf{P}\{\|P_n - P\| > \varepsilon\} &= \mathbf{E}\left[\mathbf{P}\{\|P_n - P\| > \varepsilon \mid X\}\right] \\
&\leq \mathbf{E}\left[8(n+1)e^{-n\varepsilon^2/32}\right] \\
&= 8(n+1)e^{-n\varepsilon^2/32}.
\end{aligned} \tag{2.1.5}$$

The Borel-Cantelli Lemma states that if the probability of a sequence of events is summable, that is $\sum_{n=1}^{\infty} \mathbf{P}\{E_n\} < \infty$, then

$$\limsup_n \mathbf{P}(E_n) \leq \mathbf{P}\left\{\bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} E_n\right\} = 0.$$

Since the inequality we obtain through the previous steps is exponential, the probabilities of the events $E_n = \{\|P_n - P\| > \varepsilon\}$ are summable:

$$\sum_{n=1}^{\infty} \mathbf{P}\{\|P_n - P\| > \varepsilon\} < \infty.$$

Therefore, using the Borel-Cantelli lemma we conclude that

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \to 0 \text{ with probability 1.}$$

$\square$

In chapter 4 we will elaborate further on the details required to transform this powerful theorem in a more general version.

3 Application to Estimation of Data Dimension

3.1 Chernoff-Okamoto Inequality

Let X_i be a sample from the Bernoulli distribution $\mathrm{Be}(p)$. Define $X = \sum_{i=1}^{n} X_i$, and let $\lambda = np = \mathbf{E}\, X$. Note that for $u > 0$,

$$\begin{aligned}\mathbf{E}\, e^{uX} &= \prod_i \mathbf{E}\, e^{uX_i} = ((1-p) + pe^{u})^n, \\ \mathbf{E}\, e^{-uX} &= \prod_i \mathbf{E}\, e^{-uX_i} = ((1-p) + pe^{-u})^n.\end{aligned} \tag{3.1.1}$$

By applying Markov's Inequality to e^{uX}, we can assert that

$$\begin{aligned}\mathbf{P}\{X \geq \lambda + t\} &= \mathbf{P}\{e^{uX} \geq e^{u(\lambda+t)}\} \\ &\leq e^{-u(\lambda+t)} \cdot \mathbf{E}\, e^{uX} \\ &= e^{-u(\lambda+t)} \cdot (1 - p + pe^{u})^n.\end{aligned}$$

According to Janson (2002), the right hand equation is minimized when,

$$e^u = \frac{\lambda + t}{(n - \lambda - t)} \cdot \frac{1-p}{p}.$$

Therefore, for $0 \leq t \leq n - \lambda$,

$$\mathbf{P}\{X \geq \lambda + t\} \leq \left(\frac{\lambda}{\lambda + t}\right)^{\lambda+t} \left(\frac{n-\lambda}{n-\lambda-t}\right)^{n-\lambda-t}. \tag{3.1.2}$$

However, a simpler expression is required for the following application.

Theorem 3.1. Let X be the random variable we defined at the start of this chapter. In particular, X is a random variable with the binomial distribution $\mathrm{Bi}(n,p)$ with $\lambda := np = \mathbf{E}\, X$, then for $t \geq 0$,

$$\mathbf{P}\{X \leq \lambda - t\} \leq \exp\left(-\frac{t^2}{2\lambda}\right). \tag{3.1.3}$$

Used in: Theorem 3.3

Proof. This proof was adapted from Appendix A.1.13 from Alon and Spencer (2016). The first step is to apply formula 3.1.1

$$\begin{aligned}\mathbf{P}\{X < \lambda - t\} &= \mathbf{P}\{e^{-uX} < e^{-u(\lambda-t)}\} \\ &\leq e^{u(\lambda-t)}\mathbf{E}\, e^{-uX} \\ &= e^{u(\lambda-t)}e^{u\lambda}((1-p)+pe^{-u})^n.\end{aligned}$$

Then, use the inequality $1+u \leq e^u$ to conclude,

$$(1-p)+pe^{-u} = 1+(e^{-u}-1)p < e^{p(e^{-u}-1)}$$

$$\implies ((1-p)+pe^{-u})^n \leq e^{np(e^{-u}-1)} = e^{\lambda(e^{-u}-1)}.$$

Combining everything, we obtain

$$\mathbf{P}\{X < \lambda - t\} \leq e^{\lambda(e^{-u}-1)+\lambda u - ut}$$

Now, we employ the following inequality obtained by the Taylor series expansion,

$$e^{-u} \leq 1 - u + u^2/2.$$

After expanding, this results in

$$\mathbf{P}\{X < \lambda - t\} \leq e^{\lambda u^2/2 - ut}.$$

Finally, by replacing $u = t/\lambda$ we obtain the desired result:

$$\mathbf{P}\{X < \lambda - t\} \leq e^{-t^2/2\lambda}.$$

□

3.2 The problem

The article Díaz et al. (2019) explains how we can estimate the dimension d of a manifold M embedded on a Euclidean space of dimension m, say $\mathbb{R}^m$. First, we are going to introduce the method they used, and then, we will show how the exponential inequalities can be used to prove two important results in the paper. The procedure starts with an example on a uniformly distributed sample on a d-sphere $\mathbb{S}^{d-1} \subset \mathbb{R}^d$, but will be later generalized for samples of any distribution with a density bounded away from zero.

In the first place, let $Z_1, \ldots, Z_k$ be a i.i.d. sample uniformly distributed on $\mathbb{S}^{d-1}$. Then, we have the following formula for the variance of the angles between $Z_i, Z_j, i \neq j$:

$$\beta_d := \mathbf{Var}\left(\arccos\langle Z_i, Z_j\rangle\right) = \begin{cases} \dfrac{\pi^2}{4} - 2\displaystyle\sum_{j=1}^{k}(2j-1)^{-2}, & \text{if } d = 2k+1 \text{ is odd,} \\ \dfrac{\pi^2}{12} - 2\displaystyle\sum_{j=1}^{k}(2j)^{-2}, & \text{if } d = 2k+2 \text{ is even.} \end{cases} \tag{3.2.1}$$

The previous formula for the angle variance is proven in Díaz et al. (2019). In order to give more insight on how we will be choosing an estimator $\widehat{d}$ of the dimension of the sphere, consider the following theorem.

Theorem 3.2 (Bounds for β_d)**.** For every $d > 1$,

$$\frac{1}{d} \leq \beta_d \leq \frac{1}{d-1}.$$

. $\square$

Knowing that for every $d > 1$, β_d is in the interval $[\frac{1}{d}, \frac{1}{d-1}]$, one can guess the dimension of the sphere by estimating β_d, and then, taking d from the lower bound of the interval where our estimator is. Since β_d is the variance of the angles in our sphere, our best choice for an estimator is the angle's sample variance,

$$U_k = \binom{k}{2}^{-1} \sum_{i<j\leq k} \left(\arccos\langle Z_i, Z_j\rangle - \frac{\pi^2}{2}\right)^2. \tag{3.2.2}$$

In Proposition 1. of Díaz et al. (2019) the authors prove that it's the Minimum Variance Unbiased Estimator for β_d on the unit sphere.

Furthermore, the authors also prove that there are some conditions on a manifold and on the data sampling distribution for which this result can be generalized. Let $X_1, \ldots, X_n$ be a i.i.d. sample from a random distribution P on a manifold $M \subset \mathbb{R}^m$, and let $p \in M$ denote a point on the manifold. For $C > 0 \in \mathbb{R}$, let $k = \lceil C\ln(n)\rceil$ and define $R(n) = L_{k+1}(p)$ as the distance between p and its $(k+1)$-nearest neighbor. W.L.O.G. assume that $p = 0 \in M$ and that $X_1, \ldots, X_k$ are the k-nearest neighbors of p. Additionally, for the following theorems to be true, we have the following requirements:

- The distribution of the sample has a continuous density.
- The density at every point of any neighborhood of p is positive.
- The manifold is at least twice differentiable.

The following theorem uses a special inequality from Chernoff-Okamoto, and it's crucial in the idea behind this generalization.

Theorem 3.3 (Bound k-neighbors)**.** For any sufficiently large $C > 0$, we have that, there exists n_0 such that, with probability 1, for every $n \geq n_0$,

$$R(n) \leq f_{p,P,C}(n) = O(\sqrt[d]{\ln(n)/n}), \tag{3.2.3}$$

where the function $f_{p,P,C}$ is a deterministic function which depends on p, P and C.

. □

The following theorem, although it does not require concentration inequalities, is important for connecting the idea of the previous theorem to the main frame. Let $\pi : R^m \to T_pM$ denote the orthogonal projection on the Tangent Space of M at p. Also, define $W_i := \pi(X_i)$ and then normalize,

$$Z_i := \frac{X_i}{\|X_i\|}, \quad \widehat{W_i} := \frac{W_i}{\|W_i\|}. \tag{3.2.4}$$

Theorem 3.4 (Projection Distance Bounds)**.** For any $i < j \leq n$,

(i) $\|X_i - \pi(X_i)\| = O(\|\pi(X_i)\|^2)$. (3.2.5)

(ii) $\|Z_i - \widehat{W_i}\| = O(\|\pi(X_i)\|)$. (3.2.6)

(iii) The inner products (cosine of angles) can be bounded as follows:

$$|\langle Z_i, Z_j\rangle - \langle \widehat{W_i}, \widehat{W_j}\rangle| \leq Kr, \tag{3.2.7}$$

for a constant $K \in \mathbb{R}$, whenever $r \geq \max(\|\pi(X_i)\|, \|\pi(X_j)\|)$.

. □

What follows is that if we know $W_1, \ldots, W_k$ are behaved similar to a uniformly distributed sample on the sphere $\mathbb{S}^d$, then, $Z_1, \ldots, Z_k$ (the normalized k-nearest neighbors of p) also behave like they are uniformly distributed on $\mathbb{S}^d$. The following theorem is made by combining the ideas of the previous theorems.

Theorem 3.5 (Projection's Angle Variance Statistic)**.** For $k = O(\ln n)$, let

$$V_{k,n} = \binom{k}{2}^{-1} \sum_{i<j\leq k} \left(\arccos\left\langle \widehat{W_i}, \widehat{W_j}\right\rangle - \frac{\pi^2}{2}\right)^2, \tag{3.2.8}$$

and let $U_{k,n} = U_k$ from equation 3.2.2. The following statements hold

(i) $k|U_{k,n} - V_{k,n}| \stackrel{n\to\infty}{\longrightarrow} 0$, in probability. (3.2.9)

(ii) $\mathbf{E}\,|U_{k,n} - V_{k,n}| \stackrel{n\to\infty}{\longrightarrow} 0$.

. □

This last theorem is as far as this document is planned to cover. However, the last result in the paper provides the main statement. It says that if we estimate β_d as we did with $U_{k,n}$ from 3.5, and then, extract $\widehat{d}$ from the interval where $U_{k,n}$ is located, it follows that,

Theorem 3.6 (Consistency)**.** When $n \to \infty$,

$$\mathbf{P}\{\widehat{d} \neq d\} \to 0.$$

3.3 Proofs

Proof Theorem 3.2: The even and the odd cases must be distinguished:

(1): When $d = 2k+2$ is even: In the first place, remember that,

$$\lim_{k\to\infty} \sum_{j=1}^{k} j^{-2} = \frac{\pi^2}{6}.$$

It follows from the equation 3.2.1 that

$$\begin{aligned}\beta_d &= \frac{\pi^2}{12} - 2\sum_{j=1}^{k}(2j)^{-2} = \frac{\pi^2}{12} - \frac{1}{2}\sum_{j=1}^{k} j^{-2} \\ &= \frac{1}{2}\sum_{j=k+1}^{\infty} j^{-2}.\end{aligned}$$

Since $(j^{-2})_{j\in\mathbb{N}}$ is a monotonically decreasing sequence, it follows that

$$\begin{aligned}\frac{1}{d} = \frac{1}{2k+2} &= \frac{1}{2}\int_{k+1}^{\infty} x^{-2}dx \\ \leq \quad \beta_d \quad &\leq \frac{1}{2}\int_{k+1/2}^{\infty} x^{-2}dx \\ = \frac{1}{2k+1} &= \frac{1}{d-1}.\end{aligned}$$

(2): When $d = 2k+3$ is odd: On the other hand, note that

$$\begin{aligned}\lim_{k\to\infty}\sum_{j=1}^{k}(2j-1)^{-2} &= \lim_{k\to\infty}\sum_{j=1}^{2k-1} j^{-2} - \sum_{j=1}^{k-1}(2j)^{-2} \\ &= \lim_{k\to\infty}\sum_{j=1}^{2k-1} j^{-2} - \frac{1}{4}\sum_{j=1}^{k-1} j^{-2} \\ &= \frac{\pi^2}{6} - \frac{\pi^2}{24} = \frac{\pi^2}{8}.\end{aligned}$$

Hence,

$$\beta_d = \frac{\pi^2}{4} - 2\sum_{j=1}^{k}(2j-1)^{-2}$$
$$= 2\sum_{j=k+1}^{\infty}(2j-1)^{-2}.$$

Using a similar argument we conclude that

$$\begin{aligned}\frac{1}{d} = \frac{1}{2k+1} &= 2\int_{k+1}^{\infty}(2x-1)^{-2}dx \\ \leq \quad \beta_d \quad &\leq 2\int_{k+1/2}^{\infty}(2x-1)^{-2}dx \\ = \frac{1}{2k+2} &= \frac{1}{d-1}.\end{aligned}$$

$\square$

Proof Theorem 3.3: The volume of a d-sphere of radius r is equal to:

$$v_d r^d = \frac{\pi^{d/2}}{\Gamma(\frac{n}{2}+1)} r^d.$$

Where v_d is the volume of the unit d-sphere. For the assumptions we made on P and M around $p = 0$, we can say that for any $r > 0$, there's a percent (greater than 0) of the sample that is within a range r from p. This proportion is subordinated only by the volume of a d-sphere of radius r and a constant $\alpha := \alpha(P)$ that depends on the distribution P:

$$\rho = \mathbf{P}\{X \in M : |X| < r\} \geq \alpha v_d r^d > 0.$$

We can now define a binomial process based on how many neighbors does p have within a range r. Let $N = N_r \sim \mathrm{Bi}(n, \rho)$ be the number of neighbors, using Theorem 3.1 with $\lambda = n\rho$ and $t = \frac{\lambda}{2}$ we obtain,

$$\mathbf{P}\{N \leq \lambda - t\} = \mathbf{P}\{2N \leq \lambda\} \leq \exp(-\lambda/8).$$

Since $n(\alpha v_d r^d) \leq n\rho = \lambda$, it follows that, by choosing $r(n)$ such that

$$r(n) = \left(\frac{C}{\alpha v_d} \cdot \frac{\ln n}{n}\right)^{1/d} = O(\sqrt[d]{\ln(n)/n}), \qquad (\star)$$

and thus,

$$C \ln n = n(\alpha v_d r(n)^d) \leq \lambda,$$

we obtain:

$$P\{2N \leq C \ln n\} \leq \mathbf{P}\{2N \leq \lambda\},$$

and,

$$\exp(-\lambda/8) \leq \exp\left(\frac{-C\ln n}{8}\right) = n^{-C/8}.$$

Therefore,

$$P\{2N \leq C\ln n\} \leq n^{-C/8}.$$

Finally, with this last expression we proved that if $k = \frac{C}{2}\ln n$, then the k-neighbors of p are contained in the ball of radius $r(n)$ with a probability that converges exponentially to 1. □

Proof Theorem 3.4: W.L.O.G. we assumed from the beginning that $p = 0$ by translating everything. Let $(U, x_1, \ldots, x_m)$ be a chart for $p \in M$. Since the real dimension of the manifold is d, there exists by Theorem 11.5 Tu (2011), a change of basis $(U, z_1, \ldots, z_m)$ such that T_pM is spanned by $\langle z_1, \ldots, z_d\rangle$. Let

$$\phi(z_1, \ldots, z_d) = (z_1, \ldots, z_d, F_1(z_1, \ldots, z_d), \ldots, F_t(z_1, \ldots, z_d)),$$

with $t = m - d$

$$\text{dist}(z_1, \ldots, z_d) = \sqrt{F_1^2(z) + \cdots + F_t^2(z)}.$$

$$\left\| \frac{X}{\|\pi X\|} - \frac{\pi X}{\|\pi X\|} \right\| \leq K\|\pi X\|.$$

$$\begin{aligned}
\left\| \frac{X}{\|X\|} - \frac{X}{\|\pi X\|} \right\| &= \|X\| \left| \frac{1}{\|\pi X\|} - \frac{X}{\|\pi X\|} \right| \\
&= |\|X\| - \|\pi X\|| \, \|\pi X\|^{-1} \\
&\leq \|X - \pi X\| \|\pi X\|^{-1} \\
&\leq K\|\pi X\|.
\end{aligned}$$

By plugging everything together,

$$\left\| \frac{X}{\|X\|} - \frac{\pi X}{\|\pi X\|} \right\| \leq \left\| \frac{X}{\|X\|} - \frac{X}{\|\pi X\|} \right\| + \left\| \frac{X}{\|\pi X\|} - \frac{\pi X}{\|\pi X\|} \right\| \leq 2K\|\pi X\|$$

□

Proof Theorem 3.5: □

4 Applications to graph theory

4.1 The Azuma-Hoeffding Inequality

Definition 4.1. A sequence $X_0, \ldots, X_n$ of random variables is a **martingale** if, for every $i \leq n$,

$$\mathbf{E}\left[X_{i+1} | X_i, \ldots, X_0\right] = X_i.$$

A random graph $G = G(n)$ is a graph that has n labeled vertices and produces an edge between 2 of them with a probability. Let $v_1, \ldots, v_n$ denote the vertices of G and $e_1, \ldots, e_m$ all of the $\binom{n}{2}$ potential edges that G can produce. Also, define each edge's indicator function as follows,

$$\mathbb{1}_{e_k \in G} = \begin{cases} 1, & e_k \in G \\ 0, & \text{otherwise} \end{cases}$$

An edge exposure martingale is a sequence of random variables defined as the expected value of a function $f(G)$ which depends on the information of the first j potential edges:

$$X_j = \mathbf{E}\left[f(G) \mid \mathbb{1}_{e_1 \in G}, \ldots, \mathbb{1}_{e_j \in G}\right].$$

Since all of the graph information is contained in its edges, the sequence transitions from no information: $X_0 = E(f(G))$, to the true value of the function: $X_m = f(G)$. Similarly, one can define a martingale which depends on how many vertices are revealed. The vertex exposure martingale is defined as follows,

$$X_i = \mathbf{E}\left[f(G) \mid \mathbb{1}_{\{v_k, v_j\} \in G},\ k < j \leq i\right].$$

The following inequality is to some extent an adapted version of Hoeffding inequality 2.3 for martingale random variables. If we stablish a limit for which a martingale varies from one step to another, the theorem then states that we can exponentially bound the tails of its distribution:

Theorem 4.1 (Azuma-Hoeffding inequality)**.** Let $X_0, \ldots, X_m$ be a martingale with $X_0 = 0$, and

$$|X_{i+1} - X_i| \leq 1, \quad \forall i < m. \tag{4.1.1}$$

Then, for $t > 0$,

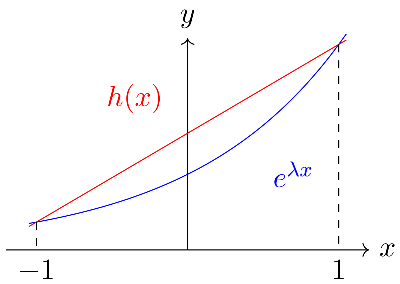
$$\mathbf{P}\{X_m > t\sqrt{m}\} < e^{-t^2/2}.$$

Proof. First, we must prove another inequality.

Lemma 4.2. Let $Y_1, \dots, Y_m$ be random variables such that $|Y_i| \le 1$ and $\mathbf{E}\, Y_i = 0$, and let $S_m = \sum_{i=1}^m Y_i$. Then, for $\lambda > 0$,

$$\mathbf{E}\left[e^{\lambda Y_i}\right] \;\le\; e^{\lambda^2/2}.$$

Proof. $\quad h(x) = \dfrac{e^{\lambda} + e^{-\lambda}}{2} + \dfrac{(e^{\lambda} - e^{-\lambda})x}{2},$



As the picture above shows, $h(x)$ is the line that passes through the points $x = -1$ and $x = 1$ in the function $e^{\lambda x}$. Since $e^{\lambda x}$ is convex ($\lambda > 0$), follows that $h(x) \ge e^{\lambda x}$ for $x \in [-1, 1]$. Thus,

$$\begin{aligned} \mathbf{E}\left[e^{\lambda Y_i}\right] &\le \mathbf{E}\left[h(Y_i)\right] \\ \text{(h is linear)} &= h(\mathbf{E}\, Y_i) = h(0) \\ &= \frac{e^{\lambda} + e^{-\lambda}}{2} = \cosh \lambda. \end{aligned}$$

Finally, $(2k)! \ge 2^k \cdot k!$, for every $k \in \mathbb{N}$. Thus,

$$\mathbf{E}\left[e^{\lambda Y_i}\right] \le \cosh \lambda \;=\; \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{(2k)!} \;\le\; \sum_{k=0}^{\infty} \frac{\lambda^{2k}}{2^k \cdot k!} \;=\; e^{\lambda^2/2}.$$

□

Now, define $Y_i = X_i - X_{i-1}$. Then, by hypothesis, $|Y_i| \le 1$ and

$$\mathbf{E}\left[Y_i | X_{i-1}, \dots, X_0\right] = \mathbf{E}\left[X_i - X_{i-1} | X_{i-1}, \dots, X_0\right] = X_i - X_i = 0.$$

Therefore, we can apply the previous inequality to assert,

$$\mathbf{E}\left[e^{\lambda Y_i} | X_{i-1}, \dots, X_0\right] \le e^{\lambda^2/2}. \qquad (\star)$$

Using the formula $E[XY] = E_X[X\,E[Y|X]]$ we assert that

$$\mathbf{E}\, e^{\lambda X_m} \;=\; \mathbf{E}\left[\prod_{i=1}^{m-1} e^{\lambda Y_i} \cdot \mathbf{E}\left[e^{\lambda Y_m} | X_{m-1}, \dots, X_0\right]\right].$$

We repeat this process n times:

$$\begin{aligned}
\mathbf{E}\, e^{\lambda X_m} &= \mathbf{E} \prod_{i=1}^{m} e^{\lambda Y_i} \\
&= \mathbf{E}\left[\prod_{i=1}^{m-1} e^{\lambda Y_i} \cdot \mathbf{E}\left[e^{\lambda Y_m} | X_{m-1}, \ldots, X_0\right]\right] && \overset{(\star)}{\leq} \mathbf{E}\left[\mathbf{E} \prod_{i=1}^{m-1} e^{\lambda Y_i}\right] e^{\lambda^2/2} \\
&= \mathbf{E}\left[\prod_{i=1}^{m-2} e^{\lambda Y_i} \cdot \mathbf{E}\left[e^{\lambda Y_{m-1}} | X_{m-2}, \ldots, X_0\right]\right] e^{\lambda^2/2} && \overset{(\star)}{\leq} \mathbf{E}\left[\mathbf{E} \prod_{i=1}^{m-2} e^{\lambda Y_i}\right] e^{2\lambda^2/2} \\
&= \qquad \vdots && \leq \qquad \vdots \\
&= \qquad \mathbf{E}\left[\mathbf{E}\left[e^{\lambda Y_1} | X_0\right]\right] e^{\lambda^2/2} && \leq \qquad e^{m\lambda^2/2}
\end{aligned} \tag{$*$}$$

At last, by setting $\lambda = t/\sqrt{m}$ we obtain,

$$\begin{aligned}
\mathbf{P}\{X_m > t\sqrt{m}\} &= \mathbf{P}\{e^{\lambda X_m} > e^{\lambda t \sqrt{m}}\} \\
\text{(Markov)} \quad &\leq \mathbf{E}\left[e^{\lambda X_m}\right] e^{-\lambda t \sqrt{m}} \\
&\overset{(*)}{\leq} e^{m\lambda^2/2} \cdot e^{-\lambda t \sqrt{m}} \\
(\lambda = t/\sqrt{m}) \quad &= e^{t^2/2} e^{-t^2} = e^{-t^2/2}.
\end{aligned} \tag{$\bullet$}$$

$\square$

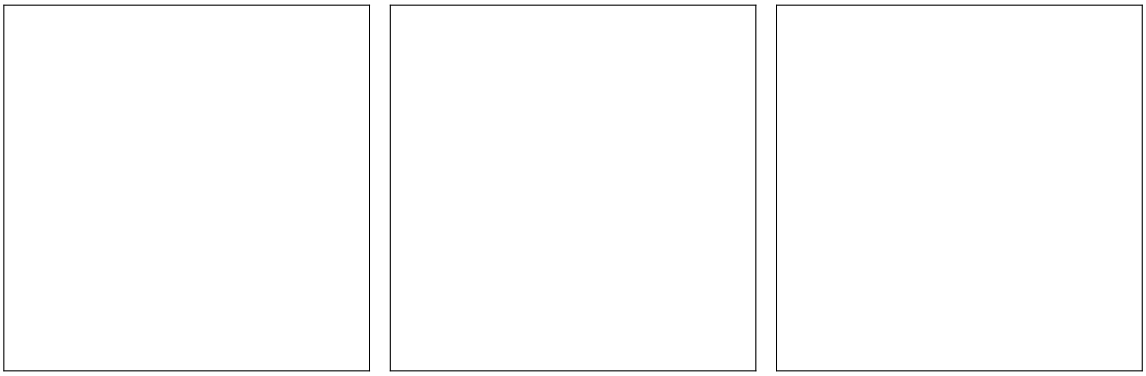
Remark. We assumed that $X_0 = 0$ to lighten the notation. However, we can remove this restriction by replacing X_m with $X_m - X_0$ in some crucial steps:

$$\begin{aligned}
& X_m - X_0 = \sum_{i=1}^{n} Y_i \\
\overset{(*)}{\Longrightarrow}\ & \mathbf{E}\, e^{\lambda (X_m - X_0)} = \mathbf{E} \prod_{i=1}^{m} e^{\lambda Y_i} \leq e^{m\lambda^2/2} \\
\overset{(\bullet)}{\Longrightarrow}\ & \mathbf{P}\{X_m - X_0 > t\sqrt{m}\} \leq e^{-t^2/2}
\end{aligned}$$

In the following section we are going to present an application of the Azuma-Hoeffding inequality in a problem involving a fast (but ineffective) approximation algorithm for the *Travelling Salesman Problem.*

4.2 An Heuristic Algorithm for the Travelling Salesman Problem

Let $X_1, \ldots, X_N$ be a sample of N uniformly distributed points in a compact square $[0, L] \times [0, L]$. The algorithm divides this square in M stripes of width L/M each. Then,



(a) Uniform sample (b) Divide in M stripes (c) Join points vertically

it connects each of the points in each of the stripes vertically and connects the top-most of one stripe with the top-most of the next one (or viceversa as the image below shows).

In the paper Gzyl et al. (1990) the authors found that the optimal number of stripes is $M^* = \lfloor 0.58N^{1/2} \rfloor$. If t_N is the TSP solution distance for our sample and d_N is the algorithm's answer with the optimal M^*, then the error is asymptotically:

$$\frac{d_N - t_N}{t_N} \approx 0.23.$$

The result that we are going to show is that d_n is very concentrated around its mean. In order to prove this, some modifications must be made to the algorithm's trajectory. Let e_N be the distance of a new trajectory that satisfies the following conditions:

- For any empty stripe in the plane we sum the length of its diagonal $\sqrt{L^2 + L^2/M^2}$ and then it skips the empty stripe.
- When there are no empty stripes, $e_N = d_N$

The probability that any given stripe is empty converges exponentially to 0,

$$\begin{aligned}(1-1/M)^N &= (1 - 0.58^{-1}N^{-1/2})^N \\ &= \left((1-1/M)^M\right)^{0.58^{-1}N^{1/2}} \\ &\sim \exp(-0.58^{-1}N^{1/2}).\end{aligned}$$

Let $\mathcal{A}_i := \sigma\{X_1, \ldots, X_i\}$ denote the sigma algebra corresponding to revealing the first i points, $\mathcal{A}_0 = \{\emptyset, [0,L]^2\}$. The expected value of the trajectory e_N given that we only know the positions of the first i points in the sample is $\mathbf{E}\,(e_N|\mathcal{A}_i)$. Define

$$Z_i = \mathbf{E}\,(e_N|\mathcal{A}_i) - \mathbf{E}\,(e_N|\mathcal{A}_{i-1}),$$

as the difference of this expectations when we reveal 1 more point. Note that since

$$\mathbf{E}\,(Z_i|\mathcal{A}_i) = \mathbf{E}\,(e_N|\mathcal{A}_i, \mathcal{A}_i) - \mathbf{E}\,(e_N|\mathcal{A}_{i-1}, A_i) = \mathbf{E}\,(e_N|\mathcal{A}_i) - \mathbf{E}\,(e_N|\mathcal{A}_i) = 0,$$

$Z_1, \ldots, Z_N$ is the difference sequence of a vertex exposure martingale.

Define $e_N^{[i]}$ as the distance of the trajectory when we remove the i-th point from the sample. Intuitively from the triangle inequality, we can obtain the following inequalities:

$$e_N^{[i]} \leq e_N \leq e_N^{[i]} + 2L/M,$$

meaning that revealing one point cannot increase more than 2 widths the distance of the trajectory. Thus,

$$\|Z_i\|_\infty = \sup_{X_1,\ldots,X_N} \|\mathbf{E}\left(e_N|\mathcal{A}_i\right) - \mathbf{E}\left(e_N|\mathcal{A}_{i-1}\right)\| \leq 2L/M.. \qquad (\star)$$

On the other hand,

$$e_N - \mathbf{E}\, e_N = \mathbf{E}\left(e_N|\mathcal{A}_N\right) - \mathbf{E}\left(e_N|\mathcal{A}_0\right) = \sum_{i=1}^{N} Z_i.$$

Therefore, by the Azuma-Hoeffding inequality,

$$\mathbf{P}\{|e_N - \mathbf{E}\, e_N| > t\} \leq 2\exp\left(\frac{-t^2}{2}\sum_{i=1}^{N}\|Z_i\|_\infty^2\right).$$

Finally,

$$\sum_{i=1}^{N}\|Z_i\|_\infty^2 \leq \frac{4NL^2}{M^2},$$

which implies that

$$\mathbf{P}\{|e_N - \mathbf{E}\, e_N| > t\} \leq 2\exp\left(\frac{-t^2}{2}\sum_{i=1}^{N}\frac{4NL^2}{M^2}\right) \sim e^{-t^2KN},$$

for some $K \in \mathbb{R}^+$.

4.3 Lipschitz Condition and Three Additional Examples

We are going to expose three examples from Alon and Spencer (2016) in order to illustrate some ideas that can be associated with the main inequality of this chapter. Furthermore, by introducing the notion of a martingale with the "Lipschitz condition", we will extend the number of cases in which we can use Azuma's inequality.

Let A, B be finite sets and let $\Omega = A^B$ be the set of all functions $g : B \to A$. Assign a probability space to each function:

$$\mathbf{P}\{g(b) = a\} = p(a, b), \quad \sum_{a\in A} p(a, b) = 1.$$

The probability for any value of $g(b)$ is independent from the value of $g(b')$ for any $b, b' \in B$. Now, fix a chain of sets

$$\emptyset = B_0 \subset B_1 \subset \ldots \subset B_m = B, \quad \mathcal{B} = \{B_i\}_{i=0}^m$$

and let $L : A^B \to \mathbb{R}$ be a functional. The martingale sequence $X_0, \ldots, X_m$ associated with the functional L and the chain $\mathcal{B}$ is defined as follows: For a fixed $h \in A^B$:

$$X_i(h) = \mathbf{E}\left[L(g) \mid g(b) = h(b), \ \forall b \in B_i\right].$$

What this means is that, given that we know the values in B_i of a function h, the martingale at the i-th step predicts the outcome of $L(h)$ based only on this information. The following definition and theorem have the purpose to make our lives easier when talking about the 'boundness' of a martingale.

Definition 4.2. A functional L is said to satisfy the Lipschitz condition if for every $i < m$: Whenever two functions g, g' differ only in $B_{i+1} - B_i$,

$$|L(g) - L(g')| \leq 1.$$

A martingale associated with a functional L has the Lipschitz condition if L has the Lipschitz condition.

In other words, if the outcome of L won't change by more than one unit if two functions g, g' vary on just one coordinate, then L has the Lipschitz condition. The following theorem will prove the connection between this condition and the hypothesis for the Azuma-Hoeffding inequality. In fact, what we're going to prove is that the Lipschitz condition is a stronger than the martingale boundness condition from formula 4.1.1.

Theorem 4.3. If a martingale associated with a functional L has the Lipschitz condition, then

$$|X_{i+1}(g) - X_i(g)| \leq 1, \quad \forall g \in A^B, \ \forall i < m.$$

Proof. The proof is adapted from Alon and Spencer (2016) chapter 7. In the original proof, the author skips many steps that I believe are not trivial. Thus, I decided to restructure the proof using the same notation they used in the source material:

Preliminaries

Our goal is to bound $|X_{i+1}(h) - X_i(h)|$, so in the first place, fix $h \in A^B$, $i \in \mathbb{N}$ and define

$$p_f^{(j)} = \mathbf{P}\{g = f \mid g(b) = h(b), \ \forall b \in B_j\}. \, \forall j \in \mathbb{N}.$$

This is the probability that $g = f$ given that g agrees on B_j (the j-th revelation) with the function h we've chosen from the beginning. Now, $\forall j \in \mathbb{N}$, define $H^{(j)} \subset A^B$ to be the set of functions f in which $f(b) = h(b)$ for every $b \in B_j$. In notation,

$$H^{(j)} = \{f \in A^B : f(b) = h(b), \ \forall b \in B_j\}.$$

Note that if $h' \notin H^{(j)}$ and $g(b) = h(b)$ for every $b \in B_j$, then it would be imposible for g to be equal to h' because there would exist $b^* \in B_j$ such that $h'(b^*) \neq h(b^*) = g(b^*)$. Thus, if $h' \notin H^{(j)}$, then $p_{h'}^{(j)} = 0$. This also implies that

$$\sum_{h' \in A^B} p_{h'}^{(j)} = \sum_{h' \in H^{(j)}} p_{h'}^{(j)} = 1.$$

Rewriting X_{i+1}

From now on, let H (without any index) refer to $H^{(i+1)} =: H$. This is notation that is used on Alon and Spencer (2016). From the definition of expected value it follows that

$$\begin{aligned} X_{i+1}(h) &= \mathbf{E}\left[L(g) \mid g(b) = h(b),\ \forall b \in B_{i+1}\right] \\ &= \sum_{h' \in A^B} L(h') \cdot \mathbf{P}\{g = h' \mid g(b) = h(b),\ \forall b \in B_{i+1}\} \\ &= \sum_{h' \in H} L(h') \cdot p_{h'}^{(i+1)}. \end{aligned}$$

Rewriting X_i

As the previous step,

$$X_i(h) = \sum_{f \in H^{(i)}} L(f) p_f^{(i)}.$$

However, we our goal is to write the sum of $X_i(h)$ only in terms of functions $h' \in H$.

For $h' \in H$, let $H[h']$ be the set of h^* such that h^*, h' that can only differ in $B_{i+1} - B_i$. In notation,

$$H[h'] = \left\{ h^* : \begin{matrix} h^*(b) = h'(b),\ \forall b \in B - B_{i+1}, \\ h^*(b) = h'(b),\ \forall b \in B_i. \end{matrix} \right\}$$

Also, define for $h^* \in H[h']$

$$q_{h^*} = \mathbf{P}\{g(b) = h^*(b),\ \forall b \in B_{i+1} \mid g(b) = h(b),\ \forall b \in B_i\}.$$

Since $B_{i+1} = (B_{i+1} - B_i) \,\dot{\cup}\, B_i$, it follows from the definition of $H[h']$ that

$$\begin{aligned} \sum_{h^* \in H[h']} q_{h^*} &= \sum_{h^* \in H[h']} \mathbf{P} \left\{ \begin{matrix} g(b) = h^*(b),\ \forall b \in B_{i+1} - B_i \\ g(b) = h^*(b),\ \forall b \in B_i \end{matrix} \,\middle|\, g(b) = h(b),\ \forall b \in B_i \right\} \\ \left({\scriptstyle h(b) = h'(b) = h^*(b) \atop b \in B_i}\right) &= \sum_{h^* \in H[h']} \mathbf{P} \left\{ \begin{matrix} g(b) = h^*(b),\ \forall b \in B_{i+1} - B_i \\ g(b) = h'(b),\ \forall b \in B_i \end{matrix} \,\middle|\, g(b) = h'(b),\ \forall b \in B_i \right\} \\ &= \sum_{h^* \in H[h']} \mathbf{P}\{g(b) = h'(b),\ \forall b \in B_{i+1} - B_i\} \\ &= 1. \end{aligned}$$

$\coprod$ is the notation I'm going to use for the disjoint union. Note that if $h'_1 \neq h'_2 \in H$, then both must differ in some $b \in B - B_{i+1}$. Thus, the following unions are disjoint

$$\begin{aligned}
\coprod_{h' \in H} \coprod_{h^* \in H[h']} \{h^*\} &= \coprod_{h' \in H} H[h'] \\
&= \coprod_{h' \in H} \left\{ h^* : \begin{array}{c} h^*(b) = h'(b),\ \forall b \in B - B_{i+1} \\ h^*(b) = h'(b) = h(b),\ \forall b \in B_i \end{array} \right\} \\
\vdots \qquad &= \coprod_{h' \in H} \left\{ h^* : \begin{array}{c} h^*(b) = h'(b),\ \forall b \in B - B_{i+1} \\ h^*(b) = h(b),\ \forall b \in B_i \end{array} \right\} \\
\downarrow \qquad &= \{h^* : h^*(b) = h(b),\ \forall b \in B_i\} \\
\coprod_{h' \in H} \coprod_{h^* \in H[h']} \{h^*\} &= H^{(i)}.
\end{aligned}$$

Then, instead of iterating over $f \in H^{(i)}$, we iterate over $h' \in H$ and $h^* \in H[h']$:

$$\begin{aligned}
\mathbf{E}\left[L(g) \mid g(b) = h(b),\ \forall b \in B_i\right] &= \sum_{f \in H^{(i)}} L(f) p_f^{(i)} \\
&= \sum_{h' \in H} \sum_{h^* \in H[h']} L(h^*) p_{h^*}^{(i)}.
\end{aligned}$$

Finally, for $h' \in H$ and $h^* \in H[h']$,

$$\begin{aligned}
&p_{h^*}^{(i)} = \\
&\mathbf{P}\{g = h^* \mid g(b) = h(b),\ \forall b \in B_i\} \\
&= \mathbf{P}\{g = h^* | g(b) = h^*(b),\ \forall b \in B_{i+1}\} \cdot \mathbf{P}\{g(b) = h^*(b),\ \forall b \in B_{i+1} | g(b) = h^*(b),\ \forall b \in B_i\} \\
&= \mathbf{P}\{g = h' | g(b) = h(b),\ \forall b \in B_{i+1}\} \cdot q_{h^*} \\
&= p_{h'}^{(i+1)} \cdot q_{h^*}.
\end{aligned}$$

$$\implies X_i(h) = \sum_{h' \in H} \sum_{h^* \in H[h']} \left[L(h^*) q_{h^*}\right] \cdot p_{h'}^{(i+1)}.$$

Bound for $|X_{i+1} - X_i|$

Combine the results from the two previous sections. For the second line, remember that $\sum_{h^* \in H[h']} q_{h^*} = 1$

$$
\begin{aligned}
|X_{i+1}(h) - X_i(h)| &= \left| \sum_{h' \in H} p_{h'}^{(i+1)} \left[L(h') - \sum_{h^* \in H[h']} L(h^*) q_{h^*} \right] \right| \\
&= \left| \sum_{h' \in H} p_{h'}^{(i+1)} \sum_{h^* \in H[h']} q_{h^*} (L(h') - L(h^*)) \right| \\
&\leq \sum_{h' \in H} p_{h'}^{(i+1)} \sum_{h^* \in H[h']} q_{h^*} |L(h') - L(h^*)|
\end{aligned}
$$

By hypothesis, $|L(h') - L(h^*)| \leq 1$. Thus,

$$
|X_{i+1}(h) - X_i(h)| \leq \sum_{h' \in H} p_{h'}^{(i+1)} \sum_{h^* \in H[h']} q_{h^*} = \sum_{h' \in H} p_{h'}^{(i+1)} = 1.
$$

□

With this theorem, we can talk with more freedom about the boundness of a martingale. The following three examples will illustrate some uses for Azuma's inequality in conjunction with the previous theorem.

Example 1

Let $g \in [n]^n$ be a random vector (uniformly chosen) with n entries, in which every entry is in $[n] = \{1, \dots n\}$. Define $L(g)$ to be the amount of number that are not included in the vector,

$$
L(g) = \#\{k \ : \ g_i \neq k, \ \forall i \in [n]\} = \sum_{k=1}^{n} \mathbb{1}_{k \notin g}.
$$

For example,

$$
L(\underset{g_1}{1}, \underset{g_2}{3}, \underset{g_3}{1}, \underset{g_4}{6}, \underset{g_5}{4}, \underset{g_6}{3}) = 2. \ \text{(because 2 and 5 are missing)}
$$

We can understand the process of choosing g as independently assigning a random number in each of its coordinates. Thus, for a number $k \in \{1, \dots, n\}$, the probability that this number is not in any of the entries of the vector is

$$
\mathbf{E}\, \mathbb{1}_{k \notin g} = \mathbf{P}\{g_i \neq k, \ \forall i\} = \prod_{i=1}^{n} P\{g_i \neq k\} = \left(1 - \tfrac{1}{n}\right)^n.
$$

Hence,

$$\mathbf{E}\,L(g) = \sum_{k=1}^{n} \mathbf{P}\{g_i \neq k,\ \forall i\} = n\big(1 - \tfrac{1}{n}\big)^n \sim \frac{n}{e}.$$

Now, define $B_i = \{1, \dots, i\}$

$$\begin{aligned}
X_0(h) &= \mathbf{E}\,L(g) \sim \tfrac{n}{e},\\
X_1(h) &= \mathbf{E}\,[L(g) \mid g_1 = h_1],\\
\vdots\ &= \ \vdots\\
X_j(h) &= \mathbf{E}\,[L(g) \mid g_i = h_i,\ \forall i \leq j],\\
\vdots\ &= \ \vdots\\
X_n(h) &= \mathbf{E}\,[L(g) \mid g_i = h_i,\ \forall i \leq n] = L(h).
\end{aligned}$$

The value of $L(g)$ can vary at most by 1 for each coordinate we reveal, so $L(g)$ has the Lipschitz condition. Then, we use theorem 4.3 and Azuma-Hoeffding inequality to conclude that

$$\mathbf{P}\{|L(g) - \tfrac{n}{e}| > t\sqrt{n}\} < 2e^{-t^2/2}.$$

Example 2

Here's a case where using theorem 4.3 will give us worse results. Let $\sigma_1, \dots, \sigma_n$ be Rademacher random variables, and $v_1, \dots, v_n$ fixed vectors in the closed unit ball. Define

$$X = \left|\sum_{i=1}^{n} \sigma_i v_i\right|.$$

The goal here is to find an exponential bound for the tail distribution of X. We create a martingale that exposes the value of σ_i one i at a time. Let $\sigma' = (\sigma'_1, \dots, \sigma'_n) \in \{-1, 1\}^n$,

$$\begin{aligned}
X_0(\sigma') &= \mathbf{E}\,|\textstyle\sum_{i=1}^n \sigma_i v_i|,\\
X_1(\sigma') &= \mathbf{E}\,[|\textstyle\sum_{i=1}^n \sigma_i v_i| \mid \sigma_1 = \sigma'_1],\\
\vdots\ &= \ \vdots\\
X_j(\sigma') &= \mathbf{E}\,[|\textstyle\sum_{i=1}^n \sigma_i v_i| \mid \sigma_i = \sigma'_i,\ \forall i \leq j],\\
\vdots\ &= \ \vdots\\
X_n(\sigma') &= \mathbf{E}\,[|\textstyle\sum_{i=1}^n \sigma_i v_i| \mid \sigma_i = \sigma'_i,\ \forall i \leq n] = X.
\end{aligned}$$

The value on one coordinate can alter X to a maximum of 2 units. Thus, we could apply theorem 4.3 to conclude that $|X_{i+1} - X_i| \leq 2$. However, note that if σ', σ^* are two n-tuple that only differ on one coordinate, follows from linearity of expectation that

$$X_i(\sigma') = \frac{1}{2}(X_{i+1}(\sigma^*) + X_{i+1}(\sigma'))$$

$$\implies X_i(\sigma') - X_{i+1}(\sigma') = \frac{1}{2}(X_{i+1}(\sigma^*) - X_{i+1}(\sigma'))$$

$$\implies |X_i(\sigma') - X_{i+1}(\sigma')| = \frac{1}{2}|X_{i+1}(\sigma^*) - X_{i+1}(\sigma')| \leq 1.$$

Thus, we can apply now Azuma's inequality and conclude the following

$$\mathbf{P}\{X - EX > t\sqrt{n}\} < e^{-t^2/2},$$

$$\mathbf{P}\{X - EX < -t\sqrt{n}\} < e^{-t^2/2}.$$

Example 3

Let ρ denote the Hamming metric in the space $\{0,1\}^n$, that is

$$\rho(x,y) = \#\{i : x_i \neq y_i\}.$$

Let $B(A,s)$ be the set $\{y \ : \ \exists x \in A, \ \rho(x) \leq s\}$. The following theorem holds,

Theorem 4.4. Let $\varepsilon, t > 0$ satisfy $\varepsilon = e^{-t^2/2}$. Then,

$$|A| \geq \varepsilon 2^n \implies |B(A, 2t\sqrt{n})| \geq (1-\varepsilon)2^n.$$

Solution: Assign a probability space to $\{0,1\}^n$ where all the points have the same probability of being chosen at random. Let $X(y) = \min_{x\in A}\rho(x,y)$, then create a martingale $X_0, \ldots, X_n$ based on the number of coordinates of $\{0,1\}^n$ exposed, that is,

$$X_j(y) = \mathbf{E}\,[\min_{x\in A}\rho(x,z) \mid z_i = y_i, \ \forall i \leq j].$$

In this case, note that if y, y' differ in just one coordinate, then

$$|X(y) - X(y')| \leq 1.$$

So we can use Azuma's inequality to conclude that

$$\mathbf{P}\{X < \mathbf{E}\,X - t\sqrt{n}\} < e^{-\lambda^2/2} = \varepsilon$$

$$\mathbf{P}\{X > \mathbf{E}\,X + t\sqrt{n}\} < e^{-\lambda^2/2} = \varepsilon.$$

Finally, since $P\{X = 0\} = |A|2^{-n} \geq \varepsilon$, follows that $\mathbf{E}\,X \leq t\sqrt{n}$. Therefore,

$$\mathbf{P}\{X > 2t\sqrt{n}\} < \varepsilon,$$

and as a consequence,

$$|B(A, 2t\sqrt{n})| = 2^n\mathbf{P}\{X > 2t\sqrt{n}\} \geq 2^n(1-\varepsilon).$$

5 Applications to Vapnik–Chervonenkis theory

5.1 Sets with Polynomial Discrimination

The version of the Glivenko-Cantelli inequality we showed on chapter 2 can be generalized in multiple ways. First, we have to make some modifications in the proof of this theorem to make it work not just on intervals of the real line. The idea is to extend this property to a specific class of sets for which the final inequality will still be satisfied:

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \leq p(n) \cdot e^{-n\varepsilon^2/32}, \text{ for a polynomial } p(n). \tag{5.1.1}$$

Remember from chapter 2 that:

- X_i is a i.i.d. sample from a probability measure P.
- $P_n(A) = n^{-1} \sum \mathbb{1}_{X_i \in A}$ is the empirical measure given by n sample points.
- σ_i is a Rademacher random variable.

In chapter 2 we assumed that P is only defined on real intervals $(-\infty, t)$. Then, in the section maximal inequality, we strategically defined $(n+1)$ different disjoint intervals when ordering the sample

$$A_0 = (-\infty, X_{(1)}],\ A_1 = (X_{(1)}, X_{(2)}],\ \ldots,\ A_{n-1} = (X_{(n-1)}, X_{(n)}],\ A_n = (X_{(n)}, \infty].$$

In each one of these intervals, we fixed a representative $t_j \in A_j$ so the function

$$P_n^\circ(B) = n^{-1} \sum_{i=1}^{n} \sigma_i \mathbb{1}_{X_i \in B},$$

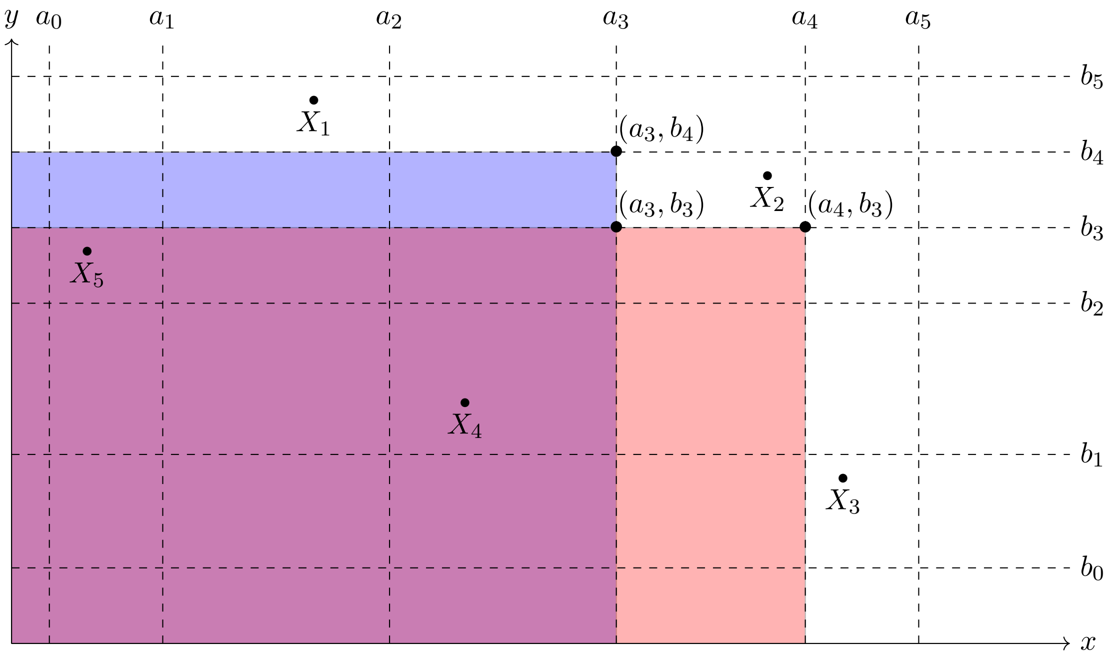
reaches its supremum in one of the sets $B_k = (-\infty, t_k)$:

$$\implies \exists k \leq n:\ \|P_n^\circ\| = |P_n^\circ(B_k)|.$$

Therefore, the $(n+1)$ term appears in the equation 2.1.3.

Quadrants in $\mathbb{R}^2$

Now, imagine that instead of $(n+1)$ intervals we take $(n+1)^2$ quadrants in the form $(-\infty, a_i) \times (-\infty, b_j) \subseteq \mathbb{R}^2$:



Let $A_{i,j} = (-\infty, a_i) \times (-\infty, b_j)$ be the quadrants described previously. In this example, we choose a_i and b_i in such way that the a_i's separate the sample horizontally and b_j vertically (similar to how we did with the t_j's in the 1-D case). Now, let $\mathscr{A}_n = \{A_{i,j}\}_{i,j \le n}$, and let $\mathscr{A}$ be the collection of all quadrants in $\mathbb{R}^2$. We will see that even though $\mathscr{A}_n \subset \mathscr{A}$ is finite, it contains all of the information of P_n°.

Let X_j^i be the i-th coordinate of the point X_j, the formula for P_n° at a point $(x, y) \in \mathbb{R}^2$ is:

$$P_n^\circ(x, y) = P_n^\circ((-\infty, x) \times (-\infty, y)) = n^{-1} \sum_{k=1}^{n} \sigma_i \mathbb{1}_{X_k^1 < x} \cdot \mathbb{1}_{X_k^2 < y}.$$

Then, because of the way we chose a_i and b_j, there exists i, j such that $x \in (a_{i-1}, a_i)$ and $y \in (b_{j-1}, b_j)$. Thus,

$$\forall k \le n : \quad \begin{array}{l} \mathbb{1}_{X_k^1 < x} = \mathbb{1}_{X_k^1 < a_i} \\ \mathbb{1}_{X_k^2 < y} = \mathbb{1}_{X_k^2 < b_j} \end{array}.$$

It follows that all the relevant information of $\mathscr{A}$ is contained in $\mathscr{A}_n$ since $P_n^\circ(x, y) = P_n^\circ(a_i, b_j) = P_n(A_{i,j})$ for some $i, j \in \mathbb{N}$. Thus, there exist $k_1, k_2 \in \mathbb{N}$ such that

$$\|P_n^\circ\|_{\mathscr{A}} = \max_{A \in \mathscr{A}_n} |P_n^\circ(A)| = |P_n(A_{k_1,k_2})|.$$

Hence,

$$\begin{aligned}\mathbf{P}\{\|P_n^\circ\|_{\mathscr{A}} > \tfrac{1}{4}\varepsilon \mid X\} &\le \sum_{i,j\le n} \mathbf{P}\{|P_n^\circ(A_{i,j})| > \tfrac{1}{4}\varepsilon \mid X\} \\ &\le (n+1)^2 \cdot \mathbf{P}\{|P_n^\circ(A_{k_1,k_2})| > \tfrac{1}{4}\varepsilon \mid X\}.\end{aligned} \tag{5.1.2}$$

The rest of the steps in the proof of the Glivenko-Cantelli theorem (2.4) never depended on the fact that we used intervals (we will elaborate further in the next section). Therefore, the formula 5.1.1, should be changed to:

$$\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \le (n+1)^2 \cdot e^{-n\varepsilon^2/32} \tag{5.1.3}$$

$$\implies \mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \xrightarrow{p} 0.$$

Note that the reason why the uniform convergence worked in the previous example, was because the geometry of the collection $\mathscr{A}$ allowed us to find a suitable sub-collection whose cardinality grows as polynomial of n. Otherwise, if we take, for instance, $\mathscr{A} = \mathcal{R}^2$ as the collection of all the open sets in $\mathbb{R}^2$, then, there are at least 2^n different sets in $\mathscr{A}$ because, since $\mathcal{R}^2$ is a metric space, we can always separate k of the sample points from the rest of the sample. Thus, the Glivenko-Cantelli inequality won't hold anymore:

$$\mathbf{P}\{\|P_n - P\|_{\mathbb{R}^2} > \varepsilon\} \le 2^n \cdot e^{-n\varepsilon^2/32} = e^{n(\log 2 - \varepsilon^2/32)}, \tag{5.1.4}$$

which diverges to ∞ when $\varepsilon \le \sqrt{\log 2^{32}}$. This will introduce us to the definition we're looking for.

Definition 5.1. A collection of sets $\mathscr{A}$ of some space S is said to have a polynomial discrimination of degree v if there exists a polynomial $p(\cdot)$ such that:

- For any given n points $X_1, \dots, X_n \in S$, there exists a sub-collection $\mathscr{A}_n$ such that for any set $A \in \mathscr{A}$, there exists $B \in \mathscr{A}_n$ that satisfies $\mathbb{1}_{X_i\in A} = \mathbb{1}_{X_i \in B}$ for every $i \le n$.
- The size of $\mathscr{A}_n$ is at most $p(n)$: $\#\mathscr{A}_n \le p(n) = O(n^v)$.

An equivalent way to express this definition is to say that for any subspace $S_n = \{X_1, \dots, X_n\} \subset S$, there are at most $p(n)$ different sets with the form $A \cap S_n$ for $A \in \mathscr{A}$:

$$\max_{X_1,\dots,X_n\in S} \#\{A \cap \{X_1, \dots, X_n\} \mid A \in \mathscr{A}\} \le p(n) \le 2^n.$$

Remark. For any collection $\mathscr{A}$ and a sample $X_1, \dots, X_n$ there exists a sub-collection $\mathscr{A}_n$ such that

$$\#\mathscr{A}_n = \#\{A \cap \{X_1, \dots, X_n\} \le 2^n.$$

Define the equivalence relationship $\simeq$ as it follows,

$$A \simeq B \iff \forall i \le n:\ \mathbb{1}_{X_i\in A} = \mathbb{1}_{X_i\in B},$$

which is in turn equivalent to

$$A \simeq B \iff \forall i \leq n: \; A \cap \{X_1, \dots, X_n\} = B \cap \{X_1, \dots, X_n\}.$$

This equivalence proves that both of the definitions are the same. Then, in order to construct $\mathscr{A}_n$ take one representative in each of the $\#\{A \cap \{X_1, \dots, X_n\}\}$ different equivalence classes $[A]_\simeq$, $A \in \mathscr{A}$.

Another important fact from the previous remark is that, for any collection $\mathscr{A}$, and any given sample $X_1, \dots, X_n$, since for every set $A \in \mathscr{A}$ there exists a set $B \in \mathscr{A}_n$ such that $\mathbb{1}_{X_i \in A} = \mathbb{1}_{X_i \in B}$, $\forall i \leq n$ and $\#\mathscr{A}_n \leq 2^n$, it follows that $\|P_n^\circ\|_{\mathscr{A}}$ exists and,

$$\exists A^\star \in \mathscr{A}_n : \sup_{A \in \mathscr{A}} \|P_n^\circ(A)\| = \max_{B \in \mathscr{A}_n} |P_n^\circ(B)| = |P_n^\circ(A^\star)|.$$

Similar to the quadrants example in the equations 5.1.2 and 5.1.3, we conclude that if $\mathscr{A}$ has a polynomial discrimination, then

$$\begin{aligned} \mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid X\} &\leq \sum_{A \in \mathscr{A}_n} \mathbf{P}\{|P_n^\circ(A^\star)| > \tfrac{1}{4}\varepsilon \mid X\} \\ &= \#\mathscr{A}_n \cdot \mathbf{P}\{|P_n^\circ(A^\star)| > \tfrac{1}{4}\varepsilon \mid X\} \\ &\leq p(n) \cdot \mathbf{P}\{|P_n^\circ(A^\star)| > \tfrac{1}{4}\varepsilon \mid X\}. \end{aligned} \tag{5.1.5}$$

$$\begin{aligned} &\implies \mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \leq p(n) \cdot e^{-n\varepsilon^2/32} \\ &\implies \mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \xrightarrow{p} 0. \end{aligned} \tag{5.1.6}$$

It's clear that $\mathcal{R}^2$ doesn't have polynomial discrimination. Another example of a class of sets without discrimination degree is the collection of closed convex sets on $\mathbb{S}^1 \subset \mathbb{R}^2$. For every of the 2^n subsets of any n points on the sphere, we can find a convex polygon that captures k of the points and excludes the rest. We are going to show how this works for $n = 5$:

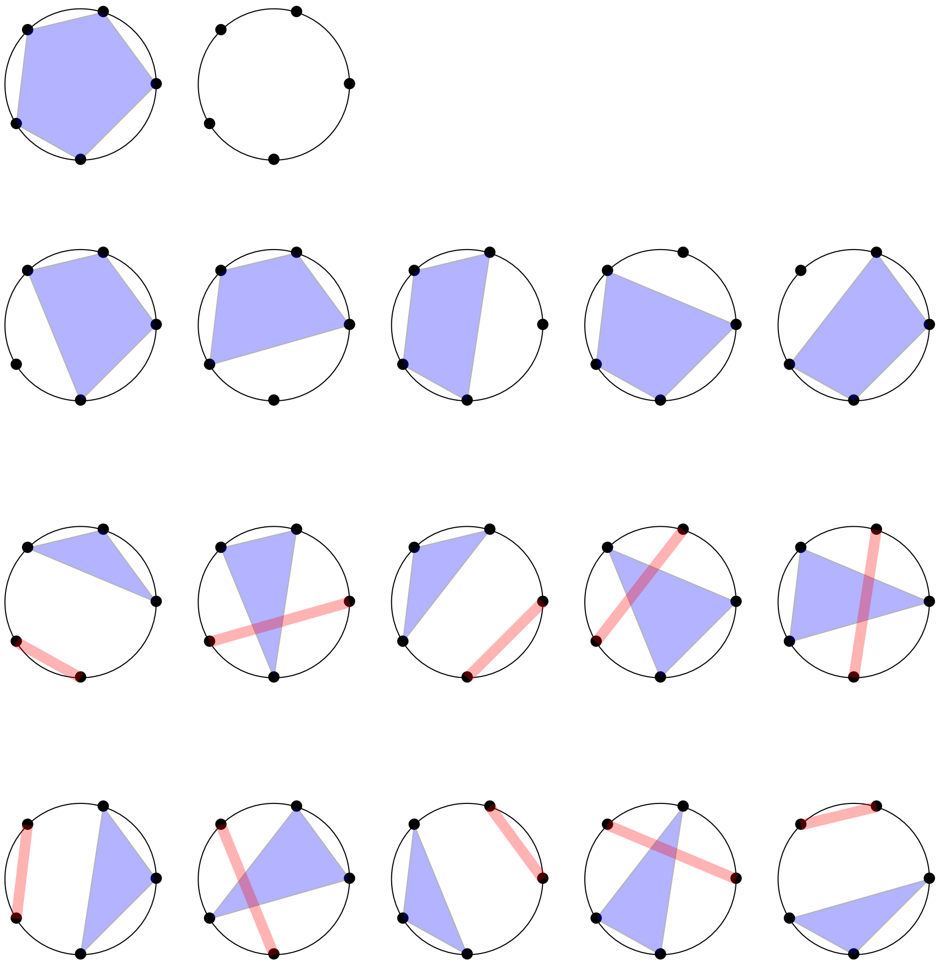


Figure 5.1: All 32 unique subsets of 5 points on $\mathbb{S}^1$

5.2 Vapnik–Chervonenkis inequality

In the previous section we conclude that the uniform law of large numbers is satisfied for collections of sets with polynomial discrimination.

Definition 5.2. Let $N_{\mathscr{A}}(X_1, \ldots, X_n)$ be the number of different sets with the form $\{X_1, \ldots, X_n\} \cap A$ for $A \in \mathscr{A}$

$$N_{\mathscr{A}} = \#\{\{X_1, \ldots, X_n\} \cap A\ ;\ A \in \mathscr{A}\}.$$

The n-th shatter coefficient of the collection $\mathscr{A}$ is the maximum of $N_{\mathscr{A}}$ over all possible points in S:

$$s(\mathscr{A}, n) = \max_{X_1, \ldots, X_n \in S} N_{\mathscr{A}}(X_1, \ldots, X_n) \leq 2^n.$$

Finally, the Vapnik–Chervonenkis dimension is defined as the largest integer k for which $s(\mathscr{A}, n) = 2^k$,

$$V_A = \operatorname*{argmax}_{k\in\mathbb{N}}\{s(\mathscr{A}, k) = 2^k\} = \operatorname*{argmin}_{k\in\mathbb{N}}\{s(\mathscr{A}, k) < 2^k\} - 1.$$

If $s(\mathscr{A}, n) = 2^n$ for every $n \in \mathbb{N}$ or equivalently if $\mathscr{A}$ doesn't have polynomial discrimination, we say that $V_A = \infty$.

Theorem 5.1 (Vapnik–Chervonenkis inequality)**.**

$$\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \le 8s(\mathscr{A}, n) \cdot e^{-n\varepsilon^2/32}.$$

Proof. Let's recapitulate everything we've done so far:

- **First Symmetrization:** Using lemma 2.5 and Chebyshev's inequality we concluded that for an identical independent copy of the empirical measure P'_n we have
$$\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \le 2\,\mathbf{P}\{\|P_n - P'_n\|_{\mathscr{A}} > \tfrac{1}{2}\varepsilon\}, \quad \text{for } n \ge \frac{8}{\varepsilon^2}.$$
- **Second Symmetrization:** We build another distribution $P_n^\circ(A) = n^{-1}\sum \sigma_i \mathbb{1}_{X_i\in A}$ and concluded from lemma 2.6 equation 2.1.2 that
$$\mathbf{P}\{\|P_n - P'_n\|_{\mathscr{A}} > \tfrac{1}{2}\varepsilon\} \le 2\,\mathbf{P}\{\|P_n^\circ\|_{\mathscr{A}} > \tfrac{1}{4}\varepsilon\}.$$
- **Maximal Inequality:** This was the step in which we had to be most careful. In the rest of the steps it never really mattered if we worked with intervals or any other class of sets on any space. In this step the task is, for any given a sample $X_1, \ldots, X_n$, to find a sub-collection $\mathscr{A}_n \subset \mathscr{A}$ such that
$$\#\mathscr{A}_n = \#\{\{X_1, \ldots, X_n\} \cap A\,;\ A \in \mathscr{A}\} = N_{\mathscr{A}}(X_1, \ldots, X_n).$$
We proved the existence of this set in the previous theorem. Then, it follows that for a given sample $X = X_1, \ldots, X_n$, the supremum of $|P_n^\circ|$ is reached in one of the sets $A^\star \in \mathscr{A}_n$. Thus,
$$\begin{aligned}\mathbf{P}\{\|P_n^\circ\|_{\mathscr{A}} > \tfrac{1}{4}\varepsilon|X\} &\le \sum_{A\in\mathscr{A}_k}\mathbf{P}\{|P_n^\circ(A)| > \tfrac{1}{4}\varepsilon|X\}\\ &\le N_{\mathscr{A}}(X)\mathbf{P}\{|P_n^\circ(A^\star)| > \tfrac{1}{4}\varepsilon|X\}\end{aligned}$$
- **Exponential Bound and integration:** After we apply Hoeffding's inequality, we obtain
$$\mathbf{P}\{\|P_n^\circ\|_{\mathscr{A}} > \tfrac{1}{4}\varepsilon|X\} \le 2N_{\mathscr{A}}(X)e^{-n\varepsilon^2/32}.$$
Finally, the result of the last expected value is
$$\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \le 8\mathbf{E}\,[N_{\mathscr{A}}(X)]e^{-n\varepsilon^2/32} \le 8s(\mathscr{A}, n) \cdot e^{-n\varepsilon^2/32}.$$

□

The middle term in the last formula is valuable to make a stronger assessment about the condition for the uniform law of large numbers. If

$$\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\} \leq 8\mathbf{E}\,[N_{\mathscr{A}}(X)]e^{-n\varepsilon^2/32},$$

According to Devroye et al. (2013), in order for $\mathbf{P}\{\|P_n - P\|_{\mathscr{A}} > \varepsilon\}$ to converge to 0 by the Borel-Cantelli theorem, the following condition must be met so the series $\sum_n 8\mathbf{E}\,[N_{\mathscr{A}}(X)]e^{-n\varepsilon^2/32}$ is summable:

$$\frac{\mathbf{E}\,[\log N_{\mathscr{A}}(X)]}{n} \to 0.$$

5.3 Estimation Error in Decision Functions

Let (X, Y) denote a pair of random variables that take values in $S \times \{0, 1\}$. The behavior of this pair can be explained by two probability functions μ, η. While μ describes the distribution of X in its space:

$$\mu(A) = \mathbf{P}\{X \in A\},$$

η describes which values of Y are more probable if $X = x$:

$$\eta(x) = \mathbf{P}\{Y = 1 | X = x\} = \mathbf{E}\,[Y | X = x].$$

A classifier or a decision function is any function tries to predict the value of Y on any given X:

$$\phi \in \mathscr{C},\ \phi : S \mapsto \{0, 1\}.$$

There's of course a probability that a classifier fails to predict correctly the value of Y. Even the best possible classifier $\phi^\star(\cdot) = \lceil 2\eta(\cdot) - 1 \rceil$ has a chance of making a mistake if $\eta(x) \neq 1$ or $\eta(x) \neq 0$. The probability of this event is called L:

$$L(\phi) = \mathbf{P}\{\phi(X) \neq Y\}.$$

The lowest possible error L^* for any classifier is called the Bayes error.

In reality, we know from little to nothing about L. We can only count on a number of observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ to decide if a classifier works. From these observations we can create an empirical function that evaluates how well a classifier fits to the observations:

$$\widehat{L}_n(\phi) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\phi(X_i) \neq Y_i}.$$

On the other hand, to optimize the computational cost, we might just consider a collection $\mathscr{C}$ of classifiers instead of all the 2^X possible functions. Let $\phi_n^\star$ be the best classifier in $\mathscr{C}$ according to $\widehat{L}_n$, that is

$$\phi_n^\star = \underset{\phi\in\mathscr{C}}{\operatorname{argmin}}\{\widehat{L}_n(\phi)\}.$$

With all the tools we've built in this chapter, we can make powerful assertions about the convergence of the error of $\phi_n^\star$.

Definition 5.3. Let $\mathscr{C}$ be a collection of decision functions $\phi : S \to \{0,1\}$. Define $\mathscr{A}$ as the following collection of sets:

$$\big\{\{\phi^{-1}(1)\times\{0\}\cup\{\phi^{-1}(0)\times\{1\}\}\big\}_{\phi\in\mathscr{C}}.$$

Define the n-shatter coefficient, and VC dimension of a classifier as

$$s(\mathscr{C},n) = s(\mathscr{A},n), \quad V_{\mathscr{C}} = V_{\mathscr{A}}.$$

Theorem 5.2. For any collection of classifiers $\mathscr{C}$ and L, $\widehat{L}_n$ as defined above,

$$\mathbf{P}\{\|\widehat{L}_n - L\|_{\mathscr{C}} > \varepsilon\} \le 8s(\mathscr{C},n)e^{-n\varepsilon^2/32}.$$

Proof. Apply theorem 5.1 on $\mathscr{A}$. □

Theorem 5.3. For the empirically selected classifier $\phi_n^\star \in \mathscr{C}$,

$$L(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi) \le 2\|\widehat{L}_n - L\|_{\mathscr{C}}.$$

Proof. Taken from Devroye et al. (2013) Lemma 8.2.

$$\begin{aligned} L(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi) &= L(\phi_n^\star) - \widehat{L}_n(\phi_n^\star) + \widehat{L}_n(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi) \\ &\le |\widehat{L}_n(\phi_n^\star) - L(\phi_n^\star)| + |\widehat{L}_n(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi)| \\ &\le \sup_{\phi\in\mathscr{C}} |\widehat{L}_n(\phi) - L(\phi)| + \sup_{\phi\in\mathscr{C}} |\widehat{L}_n(\phi) - L(\phi)| \\ &= 2\|\widehat{L}_n - L\|_{\mathscr{C}}. \end{aligned}$$

□

Therefore, we can conclude from the two previous theorems that

$$\mathbf{P}\{L(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi) > \varepsilon\} \le 8s(\mathscr{C},n)e^{-n\varepsilon^2/128}.$$

This last formula says that if the shatter coefficient is small enough, then the estimation error

$$L(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi)$$

will converge almost surely to 0. Note though that this doesn't mean that the empirical error

$$L(\phi_n^\star) - L^\star$$

will converge to 0, if the collection $\mathscr{C}$ is too small, then the approximation error

$$\inf_{\phi\in\mathscr{C}} L(\phi) - L(\phi^\star)$$

might not converge to 0 because we are under-fitting. On the other hand, if the collection $\mathscr{C}$ is too big, the approximation error will be small but, $s(\mathscr{C}, n)$ might be so big that we will have no guarantee that the estimation error will converge to 0. In conclusion, the challenge is to find a sweet spot for the size of the collection $\mathscr{C}$ so the empirical error can converge to 0:

$$\underbrace{L(\phi_n^\star) - L^\star}_{\text{emp. error}} = \left(\underbrace{L(\phi_n^\star) - \inf_{\phi\in\mathscr{C}} L(\phi)}_{\text{est. error}}\right) + \left(\underbrace{\inf_{\phi\in\mathscr{C}} L(\phi) - L^\star}_{\text{approx. error}}\right).$$

Bibliography

Noga Alon and Joel H Spencer. *The probabilistic method.* John Wiley & Sons, 2016.

Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities. In *Summer school on machine learning*, pages 208–240. Springer, 2003.

Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 2013.

Mateo Díaz, Adolfo J Quiroz, and Mauricio Velasco. Local angles and dimension estimation from data on manifolds. *Journal of Multivariate Analysis*, 173:229–247, 2019.

H Gzyl, R Jiménez, and AJ Quiroz. The physicist's approach to the travelling salesman problem—ii. *Mathematical and Computer Modelling*, 13(7):45–48, 1990.

Svante Janson. On concentration of probability. *Contemporary combinatorics*, 10(3): 1–9, 2002.

David Pollard. *Convergence of stochastic processes.* David Pollard, 1984.

Loring W Tu. Manifolds. In *An Introduction to Manifolds*, pages 47–83. Springer, 2011.