

Contents

1

1	Introduction	1	2
1.1	Basic inequalities and theorems	1	3
1.2	Why bother?	2	4
1.2.1	Coin Tossing	2	5
1.2.2	Central Limit Theorem	2	6
2	Exponential Inequalities	4	7
2.1	Which inequality is better?	6	8
2.2	Uniform Law of Large Numbers	8	9
3	Application to Estimation of Data Dimension	13	10
3.1	Chernoff-Okamoto Inequalities	13	11
3.2	Proofs	16	12
4	Applications to graph theory	19	13
4.1	The Azuma-Hoeffding Inequality	19	14
4.2	An heuristic algorithm for the Travelling Salesman Problem	19	15
5	Applications to Vapnik–Chervonenkis theory	22	16

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,

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

□

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

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

□

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 at its tails, for example,

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

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 Tossing

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

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

For a fixed $\varepsilon > 0$, we may classify a coin tossing game as rigged if, after N trials, the ratio of heads vs tails in the sample is greater than $[1 + \varepsilon : 1 - \varepsilon]$, or similarly,

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

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. \(2003\)](#)

1 Introduction

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,

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

□

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

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

□

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,

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

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

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

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

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:

$$|\mathbf{P}\{Z_N \geq t\} - \mathbf{P}\{Z \geq t\}| = O\left(\frac{1}{\sqrt{N}}\right).$$

□

Since the approximation error is greater than the bound, 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. In the following chapters, we will show some examples where better crafted strategies are needed in order to get bounds to the tails of the random variables.

2 Exponential Inequalities

94

Even if we are satisfied with the linear convergence rate provided by Chebyshev's inequality, there are simple ways to improve this bound. The following result will provide the idea from which the exponential inequalities derive

95

96

97

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

98

99

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

100

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,

101

102

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

103

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

104

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

105

106

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

107

□ 108

The following two theorems are examples on how we can obtain tighter bounds than the ones provided by Chebyshev's inequality. In particular, these theorems are derived from the idea of the previous theorem and are considered as corollaries by some authors.

109

110

111

Theorem 2.2 (Chernoff's inequality). Let $X_i \sim \text{Be}(p_i)$ 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

112

113

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

114

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

115

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

116

2 Exponential Inequalities

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, \dots, X_N be independent random variables, such that $X_i \in [a_i, b_i]$ for every $i = 1, \dots, N$. 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 hands 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 .

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

2 Exponential Inequalities

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,

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

Finally, use the inequality from Theorem 2.1 to conclude that

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

□

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

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

However, this raises the question of which of the previous inequalities is better for a given problem. In the previous case, we chose Hoeffding's inequality, but when dealing with any specific problem, one needs to determine the criteria for deciding whether it's more appropriate to use Chernoff, Hoeffding, or any other inequality. In the following section, we will try to identify situations where one of these inequalities is more suitable than the other.

2.1 Which inequality is better?

Let's start with a small improvement of the Chebyshev's bound for the one-sided tails

2 Exponential Inequalities

Theorem 2.4 (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,

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

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,

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

By linearity of expectation,

$$\mathbf{E}[(X - \mu + u)^2] = \mathbf{E}[(X - \mu)^2] + 2u \cdot \underbrace{\mathbf{E}(X - \mu)}_0 + \mathbf{E}(u^2) = \sigma^2 + u^2.$$

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

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

□

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

Corollary 2.4.1 (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 $t^2 + \sigma^2 \leq 2t^2$, or equivalently, when $\sigma^2 \leq t^2$. However, in this case both inequalities give bounds greater than 1, and thus, are useless. Therefore, we conclude that in general Chebyshev's is better for two-sided tails and Cantelli's for one-sided tails.

2.2 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:

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

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 2.5 (Glivenko-Cantelli Theorem). For P , P_n and t from above,

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

Proof. The proof, adapted from Pollard (2012), consists of 5 steps. At first instance, the author clarifies that we must establish the condition of $t \in \mathbb{Q}$ to avoid problems with measurability. The author later proves that the theorem is true for any $t \in \mathbb{R}$, 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 polynomials, and we will cover more about this in section 5.

First Symmetrization

In the first place, define P'_n as the empirical measure obtained from an independent copy of the 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 have:

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

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

For now, fix $\varepsilon > 0$, and keep in mind the values $Z = P_n - P$, $Z' = P'_n - P$, $\alpha = \frac{1}{2}\varepsilon$ and $\beta = \frac{1}{2}$.

Lemma 2.6. Let $\{Z(t)\}_{t \in T}$ and $\{Z'(t)\}_{t \in T}$ be independent stochastic processes under the same set of indices T . Also, assume that there exist $\alpha, \beta > 0$ such that

$$\mathbf{P} \left\{ \sup_{t \in T} |Z(t)| \leq \alpha \right\} \geq \beta.$$

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

$$\mathbf{P} \left\{ \sup_{t \in T} |Z(t)| > \varepsilon \right\} \leq \beta^{-1} \mathbf{P} \left\{ \sup_{t \in T} |Z(t) - Z'(t)| > \varepsilon - \alpha \right\}.$$

2 Exponential Inequalities

Proof. Since Z, Z' are independent, it follows from the hypothesis that for any index $\tau \in T$,

$$\mathbf{P}\{|Z'(\tau)| \leq \alpha | Z\} = \mathbf{P}\{|Z'(\tau)| \leq \alpha\} \geq \mathbf{P}\left\{\sup_{t \in T} |Z'(t)| \leq \alpha\right\} \geq \beta.$$

Now, fix τ such that $|Z(\tau)| > \varepsilon$ and use the previous inequality to conclude,

$$\begin{aligned} \beta \cdot \mathbf{P}\left\{\sup_{t \in T} |Z(t)| > \varepsilon\right\} &\leq \mathbf{P}\{|Z'(\tau)| \leq \alpha\} \cdot \mathbf{P}\{|Z(\tau)| > \varepsilon\} \\ &= \mathbf{P}\{|Z'(\tau)| \leq \alpha, |Z(\tau)| > \varepsilon\} \\ &\leq \mathbf{P}\{|Z(\tau) - Z'(\tau)| > \varepsilon - \alpha\} \\ &\leq \mathbf{P}\left\{\sup_{t \in T} |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}. \end{aligned}$$

□

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

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

Therefore, using the previous lemma, we conclude that

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \leq 2\mathbf{P}\{\|P_n - P'_n\| > \frac{1}{2}\varepsilon\}, \quad \text{if } n \geq 8\varepsilon^{-2}. \quad (1)$$

Second Symmetrization

The following trick will allow us to stop considering all of the $2n$ 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 2.7. 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 $Y_i = \mathbb{1}_{\{X'_i < t\}} - \mathbb{1}_{\{X_i < t\}}$, 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, from the independency of X_i from X'_i and their equality of distributions, the following symmetry holds:

$$\mathbf{P}\{Y_i = 1\} = \mathbf{P}\{Y_i = -1\} = \frac{1}{2}(1 - \mathbf{P}\{Y_i = 0\}) = \frac{1}{2}\mathbf{P}\{X_i \neq X'_i\}.$$

2 Exponential Inequalities

Since σ_i is 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,

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

□

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

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

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

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

Maximal Inequality

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

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

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

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

$$\begin{aligned}
 \mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid X\} &\leq \sum_{j=1}^n \mathbf{P}\{|P_n^\circ(j)| > \tfrac{1}{4}\varepsilon \mid X\} \\
 &\leq n \cdot \max_j \mathbf{P}\{|P_n^\circ(j)| > \tfrac{1}{4}\varepsilon \mid X\}.
 \end{aligned} \tag{3}$$

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

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

We use equation (3) to conclude

$$\mathbf{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid X\} \leq 2ne^{-n\varepsilon^2/32}.$$

Integration

something is missing before the following step,

$$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \leq 8ne^{-n\varepsilon^2/32}$$

Since the inequality is exponential, the probabilities are summable:

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

Therefore, using the Borel-Cantelli lemma we conclude that

2 Exponential Inequalities

$\mathbf{P}\{\|P_n - P\| > \varepsilon\} \rightarrow 0$ with probability 1. 267

□ 268

In chapter 4 we will elaborate further on the details required to transform this powerful theorem in a more generalized version. 269
270

3 Application to Estimation of Data Dimension

3.1 Chernoff-Okamoto Inequalities

Applying Markov's Inequality to $Y = e^{uX}$, we can assert that

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

The right hand equation is minimized when,

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

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

$$\mathbf{P}\{X \geq \lambda + t\} \leq \left(\frac{\lambda}{\lambda + t} \right)^{\lambda+t} \left(\frac{n - \lambda}{n - \lambda - t} \right)^{n-\lambda-t} \quad (3.1)$$

Theorem 3.1. Let X be random variable with the binomial distribution $\text{Bi}(n, p)$ with $\lambda := np = \mathbf{E} X$, then for $t \geq 0$,

$$\mathbf{P}\{X \geq \lambda + t\} \leq \exp \left(-\frac{t^2}{2(\lambda + t/3)} \right) \quad (3.2)$$

$$\mathbf{P}\{X \leq \lambda - t\} \leq \exp \left(-\frac{t^2}{2\lambda} \right) \quad (3.3)$$

Used in: Theorem 3.3

Proof. (TODO I've already written the proof on paper) \square

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 does 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 on any manifold.

In the first place, let Z_1, \dots, 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 $Z_i, Z_j, i \neq j$:

$$\beta_d := \mathbf{Var} (\arccos \langle Z_i, Z_j \rangle) = \begin{cases} \frac{\pi^2}{4} - 2 \sum_{j=1}^k (2j-1)^{-2}, & \text{if } d = 2k+1 \text{ is odd,} \\ \frac{\pi^2}{12} - 2 \sum_{j=1}^k (2j)^{-2}, & \text{if } d = 2k+2 \text{ is even.} \end{cases} \quad (3.4)$$

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 \hat{d} of the dimension of the sphere, consider the following theorem.

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

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

□

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,

$$U_k = \binom{k}{2}^{-1} \sum_{i < j \leq k} \left(\arccos \langle Z_i, Z_j \rangle - \frac{\pi^2}{2} \right)^2. \quad (3.5)$$

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 this result can be generalized for any manifold with samples of any distribution. Let X_1, \dots, X_n be a i.i.d. sample from a random distribution P on a manifold $M \subset \mathbb{R}^m$, and let $p \in M$ a point. 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, \dots, X_k are the k -nearest neighbors of p . Additionally, for the following theorem to be true, we require that at any neighborhood of p , the probability in that neighborhood is greater than 0.

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 $n \geq n_0$,

$$R(n) \leq f_{p,P,C}(n) = O(\sqrt[d]{\ln(n)/n}), \quad (3.6)$$

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

3 Application to Estimation of Data Dimension

325 . □

326 The following theorem, although it does not require concentration inequalities, is
 327 important for connecting the idea of the previous theorem to the main frame. Let
 328 $\pi : R^m \rightarrow T_p M$ be the orthogonal projection on the Tangent Space of M at p . Also,
 329 define $W_i := \pi(X_i)$ and then normalize,

$$330 \quad Z_i := \frac{X_i}{\|X_i\|}, \quad \widehat{W}_i := \frac{W_i}{\|W_i\|}. \quad (3.7)$$

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

$$332 \quad (i) \quad \|X_i - \pi(X_i)\| = O(\|\pi(X_i)\|^2) \quad (3.8)$$

$$333 \quad (ii) \quad \|Z_i - \widehat{W}_i\| = O(\|\pi(X_i)\|) \quad (3.9)$$

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

$$335 \quad |\langle Z_i, Z_j \rangle - \langle \widehat{W}_i, \widehat{W}_j \rangle| \leq Kr, \quad (3.10)$$

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

337 . □

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

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

$$343 \quad V_{k,n} = \binom{k}{2}^{-1} \sum_{i < j \leq k} \left(\arccos \langle \widehat{W}_i, \widehat{W}_j \rangle - \frac{\pi^2}{2} \right)^2, \quad (3.11)$$

344 and let $U_{k,n} = U_k$ from equation 3.5. The following statements hold

$$345 \quad (i) \quad k|U_{k,n} - V_{k,n}| \xrightarrow{n \rightarrow \infty} 0, \text{ in probability.} \quad (3.12)$$

$$346 \quad (ii) \quad \mathbf{E} |U_{k,n} - V_{k,n}| \xrightarrow{n \rightarrow \infty} 0.$$

347 . □

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

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

$$353 \quad \mathbf{P}\{\widehat{d} \neq d\} \rightarrow 0.$$

3.2 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,

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

It follows 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

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

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

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

Then,

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

3 Application to Estimation of Data Dimension

366 Using a similar argument we conclude that

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

368 □

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

$$370 \quad v_d r^d = \frac{\pi^{d/2}}{\Gamma(\frac{n}{2} + 1)} r^d.$$

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

$$376 \quad \rho = \mathbf{P}\{X \in M : |X| < r\} \geq \alpha v_d r^d > 0.$$

377 We can now define a binomial process based on how many neighbors does p has within
 378 a range r . Let $N = N_r \sim \text{Bi}(n, \rho)$ be the number of neighbors, using Theorem 3.1 with
 379 $\lambda = n\rho$ and $t = \frac{\lambda}{2}$ we obtain,

$$380 \quad \mathbf{P}\{N \leq \lambda - t\} = \mathbf{P}\{2N \leq \lambda\} \leq \exp(-\lambda/8).$$

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

$$382 \quad r(n) = \left(\frac{C}{\alpha v_d} \cdot \frac{\ln n}{n} \right)^{1/d} = O(\sqrt[d]{\ln(n)/n}), \quad (\star)$$

383 and thus,

$$384 \quad C \ln n = n(\alpha v_d r(n)^d) \leq \lambda,$$

385 we obtain:

$$386 \quad P\{2N \leq C \ln n\} \leq \mathbf{P}\{2N \leq \lambda\},$$

387 and,

$$388 \quad \exp(-\lambda/8) \leq \exp\left(\frac{-C \ln n}{8}\right) = n^{-C/8}.$$

389 Therefore,

$$390 \quad P\{2N \leq C \ln n\} \leq n^{-C/8}.$$

3 Application to Estimation of Data Dimension

Finally, with this last expression we proved that if $k = \frac{C}{2} \ln n$, then the k -neighbors of p 391
are contained in the ball of radius $r(n)$ with a probability that converges exponentially 392
to 1. \square 393

4 Applications to graph theory

4.1 The Azuma-Hoeffding Inequality

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

$$\mathbf{E}[X_{i+1} | X_i, \dots, X_0] = X_i$$

The random graph $G(n, p)$ has n labeled vertices and produces an edge between 2 of them with probability p . Let v_1, \dots, v_n denote the vertices and e_1, \dots, e_m the potential $\binom{n}{2}$ edges with the indicator function:

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

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:

$$X_j = \mathbf{E}[f(G) | \mathbb{1}_{e_1 \in G}, \dots, \mathbb{1}_{e_j \in G}]$$

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 it follows,

$$X_i = \mathbf{E}[f(G) | \mathbb{1}_{\{v_k, v_j\}}, k, j \leq i]$$

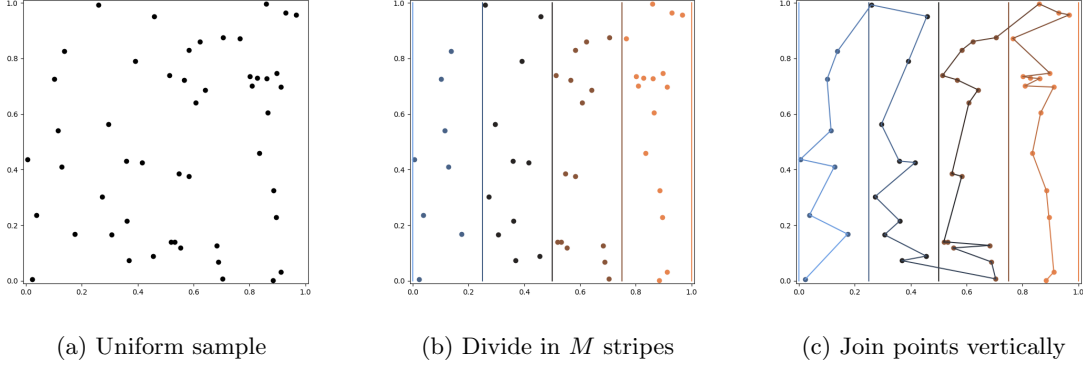
In the following section we are going to present an application of the Azuma-Hoeffding inequality to prove the convergence to the mean of a linear approximation algorithm for the *Travelling Salesman Problem*.

4.2 An heuristic algorithm for the Travelling Salesman Problem

Let X_1, \dots, 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, 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 reference [Gzyl et al. \(1990\)](#) the authors assert that by choosing a number of stripes $M^* = \lfloor 0.58N^{1/2} \rfloor$, one can achieve the best result in comparison to the real TSP

4 Applications to graph theory



solution. 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:

$$\frac{d_N - t_N}{t_N} \approx 0.23. \quad 423$$

The result that we are going to prove is that d_N converges with an exponential rate to its mean. To prove our point, we are going to modify the algorithm's trajectory as it follows. Let e_N be trajectory distance that 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 $e_N = d_N$ and the probability that any given stripe is empty converges exponentially to 0:

$$\begin{aligned} (1 - 1/M)^N &= (1 - 0.58^{-1}N^{-1/2})^N \\ &= \left((1 - 1/M)^M \right)^{0.58^{-1}N^{1/2}} \\ &\sim \exp(-0.58^{-1}N^{1/2}). \end{aligned} \quad 430$$

Let $\mathcal{A}_i := \sigma\{X_1, \dots, X_i\}$ be 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

$$Z_i = \mathbf{E}(e_N|\mathcal{A}_i) - \mathbf{E}(e_N|\mathcal{A}_{i-1}), \quad 434$$

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

$$\mathbf{E}(Z_i|\mathcal{A}_i) = \mathbf{E}(e_N|\mathcal{A}_i, \mathcal{A}_i) - \mathbf{E}(e_N|\mathcal{A}_{i-1}, \mathcal{A}_i) = \mathbf{E}(e_N|\mathcal{A}_i) - \mathbf{E}(e_N|\mathcal{A}_i) = 0, \quad 436$$

The Z_i 's form a vertex exposure martingale sequence.

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

$$e_N^{[i]} \leq e_N \leq e_N + 2L/M, \quad 440$$

4 Applications to graph theory

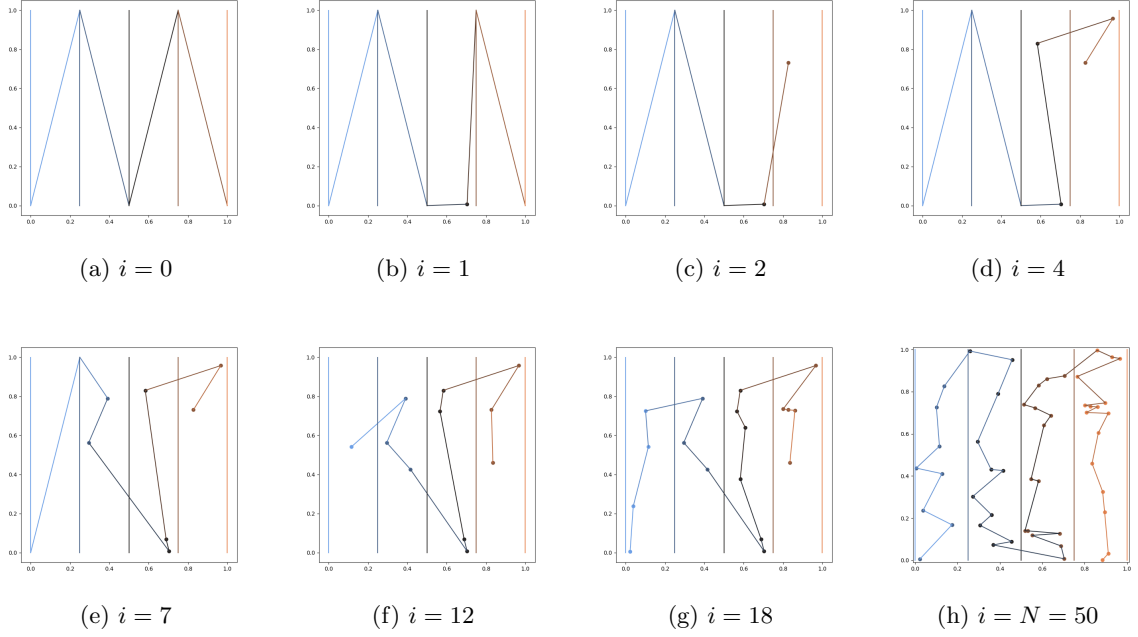


Figure 4.1: Evolution of the vertex exposure martingale

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

$$443 \quad \|Z_i\|_\infty = \sup_{X_1, \dots, X_N} \|\mathbf{E}(e_N | \mathcal{A}_i) - \mathbf{E}(e_N | \mathcal{A}_{i-1})\| \leq 2L/M.$$

444 On the other hand, by telescopic sums we obtain that

$$445 \quad e_N - Ee_N = \mathbf{E}(e_N | \mathcal{A}_N) - \mathbf{E}(e_N | \mathcal{A}_0) = \sum_{i=1}^N Z_i.$$

446 Therefore, by the Azuma-Hoeffding inequality,

$$447 \quad \mathbf{P}\{|e_N - Ee_N| > t\} \leq 2 \exp\left(\frac{-t^2}{2} \sum_{i=1}^N \|Z_i\|_\infty^2\right).$$

448 Finally,

$$449 \quad \sum_{i=1}^N \|Z_i\|_\infty^2 \leq \frac{4NL^2}{M^2},$$

450 which implies that

$$451 \quad \mathbf{P}\{|e_N - Ee_N| > t\} \leq 2 \exp\left(\frac{-t^2}{2} \sum_{i=1}^N \frac{4NL^2}{M^2}\right) \sim e^{-t^2 KN},$$

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

5 Applications to Vapnik–Chervonenkis theory

453

454

455 Bibliography

- 456 Stéphane Boucheron, Gábor Lugosi, and Olivier Bousquet. Concentration inequalities.
457 In *Summer school on machine learning*, pages 208–240. Springer, 2003.
- 458 Mateo Díaz, Adolfo J Quiroz, and Mauricio Velasco. Local angles and dimension esti-
459 mation from data on manifolds. *Journal of Multivariate Analysis*, 173:229–247, 2019.
- 460 H Gzyl, R Jiménez, and AJ Quiroz. The physicist’s approach to the travelling salesman
461 problem—ii. *Mathematical and Computer Modelling*, 13(7):45–48, 1990.
- 462 David Pollard. *Convergence of stochastic processes*. Springer Science & Business Media,
463 2012.