Universiteit Utrecht

Bachelor thesis

Concentration Bounds and Applications

Jorrit Dorrestijn

supervised by
Dr. Tobias Müller

June 6, 2016

Contents

1 Introduction

The law of large numbers states that the sample average of a sequence of independent and identically distributed random variables converges in probability towards the distribution mean as the number of samples increases. Let us state this in mathematical notation. Let $(X_i)_{i\geq 1}$ a sequence of i.i.d. random variables with sample average $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then for all $\epsilon > 0$, $\delta > 0$ there exists an integer N such that when $n \geq N$

$$\mathbb{P}\left(|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]| \geq \epsilon\right) < \delta. \tag{1}$$

Note that we have fixed the distribution a priori. What if the distribution is itself not certain? Say that the distribution is selected, unknown to us, from a collection C. We can extend Equation 1 as follows:

$$\sup_{D\in C}\left\{\mathbb{P}_D\left(|\overline{X}_n - \mathbb{E}\left[\overline{X}_n\right]| \geq \epsilon\right)\right\} < \delta$$

In this case the N provides a minimum of samples regardless the precise distribution D. In a similar fashion we can extend Equation 1 to allow different distributions for each individual X_i, $i = 1, 2, \ldots, n$. This allows us to make statements when we only have information about certain properties of the random variables involved. In general, statements providing bounds both in terms of distance and probability of samples with their expected value are called concentration bounds.

We will start with a concise overview of probability theory in Section 2. This overview includes the sigma-algebra, probability space, conditional expectation and stochastic process. In Section 3 we cover a number of concentration inequality theorems. The inequalities covered are Markov inequalities, Azuma's inequality, McDiarmid's inequality, Hoeffding's inequality and the Chernoff bound. Finally in Section 4 we will give a number of applications in the topics of statistical learning and the random geometric traveling salesman problem (TSP). Under statistical learning we cover PAC learnable, the ERM algorithm and Rademacher complexity. Under random geometric TSP we cover a result by McDiarmid and a result by Rhee and Talagrand.

2 Probability theory

In this section we provide an overview of definitions and theorems within the field of probability theory used in the rest of this paper.

A *sample space* is a set. This sample space is understood as containing the possible outcomes of an experiment. Elements of a sample space are called *samples*. A subset of a sample space is named an *event*. Typically, multiple events can occur.

2.1 Sigma-algebra

A natural thought is to group a number of events one might observe together. A formal way to describe such a family of events is the sigma-algebra.

Definition 1 (Sigma-algebra)**.** *Let Ω denote a sample space, and let $\mathcal{F}$ be a set containing subsets of Ω. Then $\mathcal{F}$ is a sigma-algebra if it satisfies*

1. *Ω is in $\mathcal{F}$*
2. *if $A \in \mathcal{F}$ then $\Omega - A \in \mathcal{F}$*
3. *if a countable number of elements $A_1, A_2, A_3, \ldots \in \mathcal{F}$ then the union $A_1 \cup A_2 \cup A_3 \cup \ldots \in \mathcal{F}$*

Elements of a sigma-algebra are called *measurable sets*. It follows that intersections of elements in a sigma-algebra are also contained in the sigma-algebra because for A, B in some sigma-algebra $\mathcal{F}$ on Ω we have that $A \cap B = \Omega - ((\Omega - A) \cup (\Omega - B))$. Based on the sigma-algebra we introduce a number of definitions.

Definition 2 (Sub-sigma-algebra)**.** *Let Ω a sample space and let $\mathcal{F}$ a sigma-algebra of Ω. Then $\mathcal{G}$ is a sub-sigma-algebra if $\mathcal{G}$ is also a sigma-algebra of Ω and $\mathcal{G} \subset \mathcal{F}$.*

Definition 3 (Generated sigma-algebra)**.** *Let Ω a sample space. Let F be a set of subsets of Ω. The sigma-algebra generated by F, notation $\sigma(F)$, is the intersection of all sigma-algebras containing F.*

Theorem 1. *Let F be a set of subsets of Ω. Then $\sigma(F)$ is a sigma-algebra.*

Proof. We will prove all properties of the sigma-algebra. Let $\mathcal{A}$ be the set of all sigma-algebras containing F. Then for all $\mathcal{F} \in \mathcal{A}$ we have that $\Omega \in \mathcal{F}$. Hence, $\Omega \in \sigma(F)$. Let $A \in \sigma(F)$. It follows that for all $\mathcal{F} \in \mathcal{A}$ we have that $A \in \mathcal{F}$. Hence, for all $\mathcal{F} \in \mathcal{A}$ we have that $\Omega - A \in \mathcal{F}$. Therefore, $\Omega - A \in \sigma(F)$. Let a countable number of elements $A_1, A_2, A_3, \ldots \in \sigma(F)$. Then each of the elements $A_1, A_2, A_3, \ldots$ is contained in all sets in $\mathcal{A}$. Hence the union of these elements is also contained in all sets in $\mathcal{A}$. Therefore the union $A_1 \cup A_2 \cup A_3 \cup \ldots$ is in $\sigma(F)$. □

Definition 4 (Borel algebra)**.** *Let $F = \{(a, b) \mid a < b \in \mathbb{R}\}$. The Borel algebra, notation $\mathcal{B}$, is the sigma-algebra generated by F.*

The following theorem lists some sets contained in the Borel algebra.

Theorem 2. *For all $a, b \in \mathbb{R}, a < b$ we have that*

- $(a, \infty) \in \mathcal{B}$
- $(\infty, a) \in \mathcal{B}$

Proof. Recall that unions of a countable number of elements in the sigma-algebra $\mathcal{B}$ are elements of $\mathcal{B}$. We have that $(a, \infty) = \cup_{n\in\mathbb{N}}(a, a + 2^n)$. Likewise, $(\infty, a) = \cup_{n\in\mathbb{N}}(a - 2^n, a)$. □

We end this subsection with the following important concept.

Definition 5 (Measurable function)**.** *Let X, Y two sets and let $\mathcal{F}_X$ the corresponding sigma-algebra of X and let $\mathcal{F}_Y$ the corresponding sigma-algebra of Y. Then the function $f : X \to Y$ is $(\mathcal{F}_X, \mathcal{F}_Y)$-measurable if the preimage of all elements in $\mathcal{F}_Y$ are contained in $\mathcal{F}_X$. That is,*

$$E \in \mathcal{F}_Y \implies f^{-1}(E) = \{x \in X \mid f(x) \in E\} \in \mathcal{F}_X.$$

2.2 Probability space

A key part of probability theory is to assign probabilities to events. For this, we define the probability measure.

Definition 6 (Probability measure)**.** *Let $\mathcal{F}$ a sigma-algebra on a sample space Ω. A function $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ is called a probability measure if it satisfies the following axioms:*

- *$\mathbb{P}$ is non-negative: for all $E \in \mathcal{F}$, $\mathbb{P}(E) \geq 0$*
- *$\mathbb{P}(\Omega) = 1$*
- *Let $E_1, E_2, E_3, \ldots \in \Omega$ a countable collection of disjoint sets. Then it must hold that $\mathbb{P}(E_1 \cup E_2 \cup E_3 \cup \ldots) = \mathbb{P}(E_1) + \mathbb{P}(E_2) + \mathbb{P}(E_3) \ldots$*

We have defined all the concepts required to define a probability space, which provides a mathematical setting to probability.

Definition 7 (Probability space)**.** *Let Ω a set, $\mathcal{F}$ a sigma-algebra on Ω and $\mathbb{P}$ a probability measure on Ω. Then the triple $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space.*

For the remainder of this probability theory section we assume the context of the probability space (Ω, $\mathcal{F}$, $\mathbb{P}$), i.e. we assume that the variables Ω, $\mathcal{F}$, $\mathbb{P}$ are defined as above.

Definition 8 (Almost surely)**.** *Within the context of a probability space, one says that an event E happens almost surely if $\mathbb{P}(E) = 1$.*

Definition 9 (Random variable)**.** *A random variable is a $(\mathcal{F}, \mathcal{B})$-measurable function $f : \Omega \to \mathbb{R}$.*

2.3 Independence

There are definitions of independence between multiple mathematical objects. We will use the following definition as basis.

Definition 10 (Independence of sigma-algebras)**.** *Let $\mathcal{F}$ and $\mathcal{G}$ be sigma-algebras. We say that $\mathcal{F}$ and $\mathcal{G}$ are independent if for all $A \in \mathcal{F}$, $B \in \mathcal{G}$ we have that*

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

To extend this definition to random variable we define a method to obtain a sigma-algebra based on random variables.

Definition 11 (Sigma-algebra generated by random variables)**.** *Let $X_1, X_2, \ldots, X_n$ random variables. We define the sigma-algebra generated by the random variables $X_1, X_2, \ldots, X_n$, notation $\sigma(X_1, X_2, \ldots, X_n)$, as follows*

$$\sigma(X_1, X_2, \ldots, X_n) = \sigma(\{X_i^{-1}(A) \mid A \in \mathcal{B}, i \in \{1, 2, \ldots, n\}\}).$$

Definition 12 (Independence of random variables)**.** *Let X and Y be random variables. We say that X and Y are independent if the generated sigma-algebras $\sigma(X)$ and $\sigma(Y)$ are independent.*

Another definition of independence which is often used is the following.

Definition 13 (Independence of random variables)**.** *Let X and Y be random variables. We say that X and Y are independent if for all $A, B \in \mathcal{B}$*

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}(X \in A)\,\mathbb{P}(X \in B).$$

We will show that the first definition is sufficient for the second definition. To avoid confusion, we will refer to the first form of independence using sigma-algebras as "independence in the first sense" and we will refer to the second form of independence using probabilities directly as "independence in the second sense".

Theorem 3. *Let X and Y be random variables. Then independence in the first sense of X and Y implies independence in the second sense of X and Y.*

Proof. Assume $\sigma(X)$ and $\sigma(Y)$ independent. Let $A, B \in \mathcal{B}$. We have that by definition

$$\mathbb{P}(X \in A, Y \in B) = \mathbb{P}\left(X^{-1}(A) \cap Y^{-1}(B)\right).$$

Also we have that $X^{-1}(A) \in \mathcal{F}$ and $Y^{-1}(B) \in \mathcal{G}$. This result together with the definition of independence in the first sense implies that

$$\mathbb{P}\left(X^{-1}(A) \cap Y^{-1}(B)\right) = \mathbb{P}\left(X^{-1}(A)\right)\mathbb{P}\left(Y^{-1}(B)\right).$$

With the identities

$$\mathbb{P}(X \in A) = \mathbb{P}\left(X^{-1}(A)\right)$$

and

$$\mathbb{P}(Y \in B) = \mathbb{P}\left(Y^{-1}(B)\right)$$

we see that $\mathbb{P}(X \in A, Y \in B) = \mathbb{P}\left(X^{-1}(A)\right)\mathbb{P}\left(Y^{-1}(B)\right)$. □

To be clear, we will use the definition of independence in the first sense in this paper from now on, unless otherwise noted.

Definition 14 (Independence of a random variable and sigma-algebra)**.** *Let X be a random variable and let $\mathcal{F}$ be a sigma-algebra. We say that X and $\mathcal{F}$ are independent if $\sigma(X)$ and $\mathcal{F}$ are independent.*

Definition 15 (Independent copy)**.** *Let X a random variable. An independent copy of X is a random variable which is independent of X and has the same distribution as X.*

2.4 Simple functions

A simple function is a function which can be written as a sum of weighted indicator functions of events.

Definition 16. *Let $n \in \mathbb{N}$, the events $E_1, E_2, \ldots, E_n \in \Sigma$ disjoint and define the corresponding constants $c_1, c_2, \ldots, c_n \in \mathbb{R}$. Then*

$$X = \sum_{i=1}^{n} c_i 1_{E_i}$$

is a simple function, where 1_{E_i} is the indicator function.

Simple functions can be used to approximate any non-negative measurable function.

Theorem 4 (Simple function approximation)**.** *Let $f : \Omega \to [0, \infty)$ a measurable function. Then a sequence of simple functions $s_1 : \Omega \to [0, \infty), s_2, s_3, \ldots$ exists such that*

1. *The sequence is non-decreasing. For all $\omega \in \Omega$ and $i \geq 1$ it holds that $s_i(\omega) \leq s_{i+1}(\omega)$.*
2. *The sequence is bounded by f. For all $\omega \in \Omega$ and $i \geq 1$ it holds that $s_i(\omega) \leq f(\omega)$.*
3. *The sequence approximates f pointwise. For all $\omega \in \Omega$ it holds that $\lim_{n\to\infty} s_n(\omega) = f(\omega)$.*

Proof. Define for all $i \geq 1$ the function $g_n : [0, \infty) \to [0, \infty)$ by

$$g_n(x) = \begin{cases} \frac{\lfloor 2^n x \rfloor}{2^n} & \text{if } x < n \\ n & \text{if } x \geq n \end{cases}$$

We will show that the sequence $(g_n \circ f)_{n\geq 1}$ is the desired sequence $(s_n)_{n\geq 1}$ by proving the three properties.

We will proof the first property with three cases. Let $n \in \mathbb{N}$. For the first case, assume that $0 \leq x < n$. In this case we have that $g_n(x) = \frac{\lfloor 2^n x \rfloor}{2^n} =$

$\frac{2\lfloor 2^n x \rfloor}{2^{n+1}} \leq \frac{\lfloor 2^{n+1} x \rfloor}{2^{n+1}} = g_{n+1}(x)$. For the second case, assume that $n \leq x < n+1$. Then $g_n(x) = n = \frac{\lfloor 2^{n+1} n \rfloor}{2^{n+1}} \leq \frac{\lfloor 2^{n+1} x \rfloor}{2^{n+1}} = g_{n+1}(x)$. For the third case, assume that $n+1 \leq x$. Then $g_n(x) = n \leq n+1 = g_{n+1}(x)$.

For the second property, observe that for all $x \geq 0$ and $n \geq 1$

$$g_n(x) = \begin{cases} \frac{\lfloor 2^n x \rfloor}{2^n} & \text{if } x < n \\ n & \text{if } x \geq n \end{cases} \leq \begin{cases} \frac{2^n x}{2^n} & \text{if } x < n \\ x & \text{if } x \geq n \end{cases} = x.$$

Hence we have that $(g_i \circ f)(x) \leq f(x)$.

For the third property, we have that for all $x \geq 0$

$$\lim_{n\to\infty} g_n(x) = \lim_{n\to\infty} \frac{\lfloor 2^n x \rfloor}{2^n} = \lim_{n\to\infty} x = x.$$

For all $\omega \in \Omega$, set $x = f(\omega)$ to obtain

$$\lim_{n\to\infty} s_n(\omega) = \lim_{n\to\infty} g_n(f(\omega)) = \lim_{n\to\infty} g_n(x) = x = f(\omega).$$

□

2.5 Convergence

An important result in probability theory is the dominated convergence theorem.

Definition 17 (Dominated convergence theorem)**.** *Let Y a random variable such that $\mathbb{E}[|Y|] < \infty$ and let $X_1, X_2, \ldots$ a sequence of random variables such that for all $n = 1, 2, \ldots$ we have that $|X_n| \leq Y$. Then if for some random variable X it holds that for all $\omega \in \Omega$ we have that $\lim_{n\to\infty} X_n(\omega) = X(\omega)$ almost surely, then $\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X]$.*

We do not proof this theorem here. Instead, we refer to [5] which provides an elementary proof for the continuous random variable case.

2.6 Conditional expectation

An important concept for the theorems outlined in this paper is conditional expectation. A general definition is the following.

Definition 18 (Conditional expectation)**.** *Let X a random variable such that $\mathbb{E}[X^2] < \infty$ and let $\mathcal{G}$ a sub-sigma-algebra of $\mathcal{F}$. Then the random variable $Y : \Omega \to \mathbb{R}$ is a conditional expectation of X given $\mathcal{G}$, denoted by $\mathbb{E}[X \mid \mathcal{G}]$, if*

- *Y is $(\mathcal{G}, \mathcal{B})$-measurable*
- *$\mathbb{E}[Y 1_A] = \mathbb{E}[X 1_A]$ for all $A \in \mathcal{G}$.*

This definition is not constructive. However, it is possible to prove that conditional expectation is unique in the almost surely sense. To provide this proof, we start with a lemma.

Lemma 5. *Let $\mathcal{G}$ a sigma-algebra. Let X and X' $(\mathcal{G}, \mathcal{B})$-measurable random variables, and let $\epsilon \in \mathbb{R}$. Then $(X - X')^{-1}[(\epsilon, \infty)] \in \mathcal{G}$.*

Proof. Let $\epsilon \in \mathbb{R}$ and let $\omega \in \Omega$. Note that $X(\omega) > X'(\omega) + \epsilon$ is equivalent with the existence of a $q \in \mathbb{Q}$ such that $X(\omega) > q > X'(\omega) + \epsilon$. Therefore we have that

$$\{\omega \in \Omega \mid X(\omega) > X'(\omega) + \epsilon\} = \cup_{q \in \mathbb{Q}} \{\omega \in \Omega \mid X(\omega) > q\} \cap \{\omega \in \Omega \mid X'(\omega) < q - \epsilon\}.$$

Rewriting with the preimage gives

$$(X - X')^{-1}[(\epsilon, \infty)] = \cup_{q \in \mathbb{Q}} X^{-1}[(q, \infty)] \cap X'^{-1}[(q - \epsilon, \infty)].$$

Because X and X' are $(\mathcal{G}, \mathcal{B})$-measurable we have that for all $q \in \mathbb{Q}$ it holds that $X^{-1}[(q, \infty)] \in \mathcal{G}$ and $X'^{-1}[(q - \epsilon, \infty)] \in \mathcal{G}$. Because $\mathcal{G}$ is a sigma-algebra we have that the union of countable elements in $\mathcal{G}$ is in $\mathcal{G}$. Hence, $(X - X')^{-1}[(\epsilon, \infty)] \in \mathcal{G}$. □

Theorem 6 (Uniqueness)**.** *Let X a random variable such that $\mathbb{E}[X^2] < \infty$ and let $\mathcal{G}$ a sub-sigma-algebra of $\mathcal{F}$. Define Y and Y' both conditional expectations of X given $\mathcal{G}$. Then*

$$\mathbb{P}(Y = Y') = 1.$$

Proof. Let $\epsilon > 0$. Define the event $E_\epsilon = (Y - Y')^{-1}[(\epsilon, \infty)] = \{\omega \in \Omega \mid Y(\omega) - Y'(\omega) > \epsilon\}$. By Lemma 5 we have that $E_\epsilon \in \mathcal{G}$. From the definition of conditional expectation it follows that

$$\mathbb{E}[Y 1_{E_\epsilon}] = \mathbb{E}[Y' 1_{E_\epsilon}] = \mathbb{E}[X 1_{E_\epsilon}].$$

It follows that

$$\mathbb{E}[(Y - Y') 1_{E_\epsilon}] = 0.$$

However, when $\omega \in E_\epsilon$ it holds that $Y(\omega) - Y'(\omega) > \epsilon > 0$. So it follows that

$$0 = \mathbb{E}[(Y - Y') 1_{E_\epsilon}] > \mathbb{E}[\epsilon 1_{E_\epsilon}].$$

Given that 1_{E_ϵ} only takes outcomes 0 and 1, it holds that $\mathbb{P}(1_{E_\epsilon} = 1) = 0$. This is by definition equal to $\mathbb{P}(Y - Y' > \epsilon) = 0$.

Define for $n \in \mathbb{N}$ $\epsilon(n) = 2^{-n}$, then define

$$E = \cup_{n \in \mathbb{N}} E_{e(n)}.$$

This gives $\mathbb{P}(1_E = 1) = 0$ which is equivalent to $0 = \mathbb{P}(Y - Y' > 0) = \mathbb{P}(Y > Y')$. Therefore $\mathbb{P}(Y \leq Y') = 1$. If we repeat the proof with the inequality signs reversed, we find that $\mathbb{P}(Y \geq Y') = 1$. We conclude that $\mathbb{P}(Y = Y') = 1$. □

From now on, we will write *the* conditional expectation instead of *a* conditional expectation, as the conditional expectation is unique in the almost surely sense.

We can condition on random variables as well by using the generated sigma algebra. This is captured in the following definition.

Definition 19 (Conditional expectation on random variables)**.** *Let Y, X_1, X_2, …, X_n random variables. Then define*

$$\mathbb{E}\left[Y \mid X_1, X_2, \ldots, X_n\right] = \mathbb{E}\left[Y \mid \sigma(X_1, X_2, \ldots, X_n)\right].$$

We list some useful properties of the conditional expectation below.

Theorem 7 (Law of total expectation)**.** *Let $\mathcal{G}$ sub-sigma-algebras of $\mathcal{F}$, and let X a random variable. Then*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right]\right] = \mathbb{E}\left[X\right].$$

Proof. From the definition of conditional expectation, we obtain that for all $A \in \mathcal{G}$

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}\right] 1_A\right] = \mathbb{E}\left[X 1_A\right].$$

If we set $A = \Omega$, then 1_A becomes the constant function $x \mapsto 1$ and the result follows. □

Theorem 8 (Tower property)**.** *Let $\mathcal{G}_1, \mathcal{G}_2$ sub-sigma-algebras of $\mathcal{F}$ such that $\mathcal{G}_1 \subset \mathcal{G}_2$ and let X a random variable. Then we have almost surely that*

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right] = \mathbb{E}\left[X \mid \mathcal{G}_1\right].$$

Proof. We show that the conditional expectation properties for $\mathbb{E}\left[X \mid \mathcal{G}_1\right]$ also hold for $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right]$. Then by uniqueness of conditional expectation we have proven the theorem.

By definition $\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right]$ is $(\mathcal{G}_1, \mathcal{B})$-measurable.

We have that for all $A \in \mathcal{G}_1$

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right] 1_A\right] = \mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] 1_A\right].$$

We also have that for all $A \in \mathcal{G}_2$

$$\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] 1_A\right] = \mathbb{E}\left[X 1_A\right].$$

Because $\mathcal{G}_1 \subset \mathcal{G}_2$ we have therefore that for all $A \in \mathcal{G}_1$

$$\mathbb{E}\left[\mathbb{E}\left[\mathbb{E}\left[X \mid \mathcal{G}_2\right] \mid \mathcal{G}_1\right] 1_A\right] = \mathbb{E}\left[X 1_A\right].$$

□

Theorem 9. *Let $\mathcal{G}$ a sigma-algebra, and let X, Y random variables. If X is $(\mathcal{G}, \mathcal{B})$-measurable, then $\mathbb{E}\left[XY \mid \mathcal{G}\right] = X\mathbb{E}\left[Y \mid \mathcal{G}\right]$.*

Proof. It suffices to show that for all $A \in \mathcal{G}$ that

$$\mathbb{E}\left[XY 1_A\right] = \mathbb{E}\left[X\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right].$$

We will first show that this proof holds for $X = 1_E$ where $E \in \mathcal{G}$. We have for all $A \in \mathcal{G}$ that

$$\mathbb{E}\left[XY 1_A\right] = \mathbb{E}\left[Y 1_{A \cap E}\right] = \mathbb{E}\left[\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_{A \cap E}\right] = \mathbb{E}\left[X\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right].$$

Let $n \in \mathbb{N}$ and define $E_1, E_2, \ldots, E_n \in \mathcal{G}$ disjoint and $a_1, a_2, \ldots, a_n \in \mathbb{R}$. We can extend this proof to all simple functions $X = \sum_{i=1}^{n} a_i 1_{E_i}$, by using linearity as follows:

$$\begin{aligned}
&\mathbb{E}\left[XY1_A\right] \\
&\qquad = \mathbb{E}\left[\sum_{i=1}^{n} a_i 1_{E_i} Y 1_A\right] \\
&\qquad = \sum_{i=1}^{n} a_i \mathbb{E}\left[1_{E_i} Y 1_A\right] \\
&\qquad = \sum_{i=1}^{n} a_i \mathbb{E}\left[1_{E_i} \mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right] \\
&\qquad = \mathbb{E}\left[\sum_{i=1}^{n} a_i 1_{E_i} \mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right] \\
&\qquad = \mathbb{E}\left[X\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right].
\end{aligned}$$

We now prove it for the class of non-negative X. Because of Theorem 4, there exists a non-decreasing sequence $X_1, X_2, \ldots$ of simple functions of the form $\Omega \to [0, \infty)$ such that for all $x \in \Omega$ we have that $X_n(x) \to X(x)$, and furthermore $X_n(x) \leq X(x)$ for all $n \in \mathbb{N}$. Therefore we have that $|X_n Y 1_A| \leq |XY1_A|$. This allows us to apply Theorem 17 (dominated convergence theorem), such that for all $A \in \mathcal{G}$

$$\mathbb{E}\left[XY1_A\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n Y 1_A\right] = \lim_{n\to\infty} \mathbb{E}\left[X_n \mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right] = \mathbb{E}\left[X\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right].$$

Finally, we can remove the non-negative restriction on X by taking positive and negative parts. Define $X^+ : \Omega \to \mathbb{R}, x \mapsto \max(X(x), 0)$ and $X^- : \Omega \to \mathbb{R}, x \mapsto -\min(X(x), 0)$. We can write $X = X^+ - X^-$. Both of X^+ and X^- are non-negative. Define $-E = \{-x \mid x \in E\}$ for some set $E \subset \mathbb{R}$. Because

$$(X^+)^{-1}(E) = \begin{cases} X^{-1}(E \cup (-\infty, 0]) & \text{if } 0 \in E \\ X^{-1}(E \cap (0, \infty)) & \text{otherwise} \end{cases}$$

$$(X^-)^{-1}(E) = \begin{cases} X^{-1}(-E \cup [0, \infty)) & \text{if } 0 \in E \\ X^{-1}(-E \cap (-\infty, 0)) & \text{otherwise} \end{cases}$$

it holds that X^+ and X^- are $(\mathcal{G}, \mathcal{B})$-measurable. Therefore we have for all $A \in \mathcal{G}$ that

$$\begin{aligned}
&\mathbb{E}\left[XY1_A\right] \\
&\qquad = \mathbb{E}\left[(X^+ - X^-)Y1_A\right] \\
&\qquad = \mathbb{E}\left[X^+\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right] - \mathbb{E}\left[X^-\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right] \\
&\qquad = \mathbb{E}\left[X\mathbb{E}\left[Y \mid \mathcal{G}\right] 1_A\right].
\end{aligned}$$

□

Theorem 10 (Pulling out independent factors)**.** *Let $\mathcal{F}$ a sigma-algebra. Let X a random variable independent of $\mathcal{F}$. Then*

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[X].$$

Proof. We will proof that $\mathbb{E}[X]$ is a conditional expectation of $\mathbb{E}[X \mid \mathcal{F}]$. Because $\mathbb{E}[X]$ is a constant, it is $(\mathcal{F}, \mathcal{B})$-measurable. For the second property, we have to prove that

$$\mathbb{E}[\mathbb{E}[X] 1_A] = \mathbb{E}[X 1_A].$$

Observe that $\sigma(1_A) = \{A, \Omega - A, \emptyset, \Omega\}$. Because $\mathcal{F}$ is a sigma-algebra containing A, we have that $\{A, \Omega - A, \emptyset, \Omega\} \subset \mathcal{F}$. From the definition of independence between $\mathcal{F}$ and X it follows that X and 1_A are independent. So we have that

$$\mathbb{E}[X 1_A] = \mathbb{E}[X]\,\mathbb{E}[1_A] = \mathbb{E}[\mathbb{E}[X] 1_A].$$

□

Theorem 11. *Let $\mathcal{F}$ a sigma-algebra and let $X_1, X_2, \ldots, X_n$ random variables. Let X_1' an independent copy of X_1 and let further X_1 and X_1' independent of all $X_2, X_3, \ldots, X_n$ and of $\mathcal{F}$. Let $f : \Omega^n \to \mathbb{R}$ an arbitrary function. For convenience, define $X = (X_1, X_2, \ldots, X_n)$ and $X' = (X_1', X_2, \ldots, X_n)$. Then almost surely*

$$\mathbb{E}[f(X) \mid \mathcal{F}] = \mathbb{E}[f(X') \mid \mathcal{F}].$$

Proof. We will show that $\mathbb{E}[f(X') \mid \mathcal{F}]$ is a conditional expectation of $\mathbb{E}[f(X) \mid \mathcal{F}]$. By definition, $\mathbb{E}[f(X') \mid \mathcal{F}]$ is $(\mathcal{F}, \mathcal{B})$-measurable. Therefore we are left to prove that for all $A \in \mathcal{F}$ it holds that

$$\mathbb{E}[\mathbb{E}[f(X') \mid \mathcal{F}] 1_A] = \mathbb{E}[f(X) 1_A].$$

By definition of $\mathbb{E}[f(X') \mid \mathcal{F}]$ this is equivalent to

$$\mathbb{E}[f(X') 1_A] = \mathbb{E}[f(X) 1_A].$$

Similar to the proof of Theorem 10, we have that X_1 and 1_A are independent. Analogously we have that X_1' and 1_A are independent. Therefore we have that both X_1 and X_1' are independent of all other random variables in the expectation, and as X_1' is an independent copy of X_1, it must be the case that the expectations are equal. □

2.7 Stochastic process

In this subsection we will provide concepts which allow us to describe experiments in which knowledge about particular random variables might change over time. The first thing to define, then, is a set of multiple random variables.

Definition 20 (Stochastic process)**.** *A stochastic process $(X_i)_{i \in I}$ is a collection of random variables for some index set I.*

Here I could be seen as the index set of time moments. To represent a changing state of knowledge over time, we define filtration.

Definition 21 (Filtration)**.** *A filtration is a collection of sigma-algebras $(\mathcal{F}_i)_{i\in I}$ for some totally ordered index set I such that if $i, j \in I$, $i < j$ then $\mathcal{F}_i \subset \mathcal{F}_j$.*

The limited knowledge of a filtration is linked to that of a random variable with the following definition.

Definition 22 (Adapted)**.** *Let I an index set. A stochastic process $(X)_{i\in I}$ is adapted to a filtration $(\mathcal{F}_i)_{i\in I}$ if X_i is $(\mathcal{F}_i, \mathcal{B})$-measurable for all $i \in I$.*

Given a stochastic process, it can be useful to define a filtration which is adapted to it. Such a filtration is the *natural filtration*.

Definition 23 (Natural filtration)**.** *Let $X = (X_1, X_2, X_3, \ldots)$ a stochastic process. Then the natural filtration of X is defined by the filtration $(\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \ldots)$ where*

$$\mathcal{N}_i = \sigma(X_1, X_2, \ldots, X_i).$$

We show that this definition is well-defined.

Theorem 12. *Let $X = (X_1, X_2, X_3, \ldots)$ a stochastic process, and let $\mathcal{N} = (\mathcal{N}_1, \mathcal{N}_2, \mathcal{N}_3, \ldots)$ the natural filtration of X. Then X is adapted to $\mathcal{N}$.*

Proof. This follows directly from the definition. □

We can use conditional expectation to define a martingale. A Martingale defines a stochastic process such that given the knowledge at time i, the expectation of the next random variable is equal to the one of time i.

Definition 24 (Martingale)**.** *Let I a totally ordered index set. Let $(X)_{i\in I}$ a stochastic process adapted to a filtration $(\mathcal{G}_i)_{i\in I}$ such that $\mathbb{E}[|Y_i|] < \infty$ for all $i \in I$. Then $(X)_{i\in I}$ is called a martingale if*

$$\mathbb{E}[X_j \mid \mathcal{G}_i] = X_i$$

for all $j > i$.

3 Concentration inequalities

In this section we provide a number of concentration inequalities.

3.1 Markov inequalities

Theorem 13 (Extended Markov inequality)**.** *Let X be a random variable and let f a strictly increasing real-valued positive function such that $\mathbb{E}[f(X)] < \infty$. Then for any $\epsilon \in \mathbb{R}$ it holds that*

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

Proof. Because f is strictly increasing the events $f(X) \geq f(\epsilon)$ and $X \geq \epsilon$ are equivalent for any ϵ. In other words

$$\mathbb{P}\left(f(X) \geq f(\epsilon)\right) = \mathbb{P}\left(X \geq \epsilon\right).$$

The result follows by the Markov inequality on $f(X)$.

□

Corollary 13.1 (Exponential Markov inequality)**.** *Let X be a random variable. Then for any $\lambda > 0$, $\epsilon \in \mathbb{R}$*

$$\mathbb{P}\left(X \geq \epsilon\right) \leq e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda X}\right]$$

Proof. Consider the function $f(x) = e^{\lambda x}$, where $\lambda > 0$. This function is non-negative and strictly increasing. Therefore we can apply this function to the extended Markov inequality to obtain for a random variable X:

$$\mathbb{P}\left(X \geq \epsilon\right) \leq \frac{\mathbb{E}\left[e^{\lambda X}\right]}{e^{\lambda\epsilon}} = e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda X}\right].$$

□

3.2 McDiarmid's inequality

The following lemma from [6] will be used to prove McDiarmid's inequality.

Lemma 14. *Let X be a random variable for which $\mathbb{E}\left[X\right] = 0$ and $a \leq X \leq b$ for some constants $a, b \in \mathbb{R}$. Then for any $\lambda > 0$*

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

Proof. Let $\lambda > 0$. We start by observing that $x \mapsto e^{\lambda x}$ is a convex function. Considering the points a and b, we have by definition of convex that for any $t \in [0, 1]$

$$e^{\lambda(ta+(1-t)b)} \leq te^{\lambda a} + (1-t)e^{\lambda b}.$$

Define $x = ta + (1-t)b$. Then $t = \frac{b-x}{b-a}$ for $x \in [a, b]$. This gives

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

hence we have

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

Let $\hat{\lambda} = \lambda(b-a)$ and $p = \frac{-a}{b-a}$. It follows that $1 - p = \frac{b}{b-a}$. Because $\mathbb{E}\left[X\right] = 0$, it holds that $a \leq 0 \leq b$. Hence $p \geq 0$. Likewise, $1 - p \geq 0$. Combining gives $0 \leq p \leq 1$. Therefore we have that

$$\begin{aligned}
&\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} \\
&\qquad = (1-p)e^{\lambda a} + pe^{\lambda b} \\
&\qquad = (1-p)e^{-p\lambda(b-a)} + pe^{(1-p)\lambda(b-a)} \\
&\qquad = (1-p)e^{-p\hat{\lambda}} + pe^{(1-p)\hat{\lambda}} \\
&\qquad = e^{-p\hat{\lambda}}\left(1-p+pe^{\hat{\lambda}}\right) \\
&\qquad = \exp\left(-p\hat{\lambda} + \log\left(1-p+pe^{\hat{\lambda}}\right)\right).
\end{aligned}$$

Define $f : \mathbb{R} \to \mathbb{R}$ by $f(x) = -px + \log(1-p+pe^x)$. Observe that $\mathbb{E}\left[e^{\lambda X}\right] \leq e^{f(\hat{\lambda})} = \exp\left(-p\hat{\lambda} + \log\left(1-p+pe^{\hat{\lambda}}\right)\right)$. We will finish the proof by bounding $f(\hat{\lambda})$. By Taylor's theorem we can write f as the sum of a linear approximation of f at 0 and a remainder:

$$f(\hat{\lambda}) = f(0) + f'(0)\hat{\lambda} + \frac{f''(\epsilon)}{2}\hat{\lambda}^2 \tag{2}$$

for some $0 \leq \epsilon \leq \hat{\lambda}$. We have

$$f'(x) = -p + \frac{pe^x}{1-p+pe^x}$$

and

$$f''(x) = \frac{(1-p)pe^x}{(1-p+pe^x)^2}.$$

For f and f' we have $f(0) = f'(0) = 0$. Plugging this in equation 2 we obtain

$$f(\hat{\lambda}) = \frac{f''(\epsilon)}{2}\hat{\lambda}^2.$$

We provide an upper bound for f''. First we show that f'' is a decreasing function on the domain $[0, \infty)$. The derivative of the numerator is $(1-p)pe^x$. The derivative of the denominator is $2pe^x(p(e^x-1)+1)$. Given that $0 \leq p \leq 1$ we see that for $x \geq 0$ the derivative of the denominator is larger or equal to the derivative of the numerator. As a result f'' is decreasing on $[0, \infty)$. Furthermore, we show that $f''(0) \leq \frac{1}{4}$:

$$f''(0) = \frac{(1-p)p}{(1-p+p)^2} = (1-p)p \leq \frac{1}{4}.$$

Combining these two results implies that $f''(\hat{\lambda}) \leq \frac{1}{4}$ for all $\hat{\lambda} > 0$.

So

$$f(\hat{\lambda}) \leq \frac{f''(\hat{\lambda})}{2}\hat{\lambda}^2 \leq \frac{1}{8}\hat{\lambda}^2 = \frac{\lambda^2(b-a)^2}{8}.$$

□

We show in the following corollary that Lemma 14 also holds for conditional expectations.

Corollary 14.1. *Let $\mathcal{F}$ a sigma-algebra. Let X be a random variable for which $\mathbb{E}[X \mid \mathcal{F}] = 0$ and $a \leq X \leq b$ for some constants $a, b \in \mathbb{R}$. Then for any $\lambda > 0$*

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

Proof. The proof is the same as the proof of Lemma 14, except that expectations are replaced by expectations conditioned on $\mathcal{F}$. Namely we have that

$$\mathbb{E}\left[e^{\lambda X} \mid \mathcal{F}\right] \leq \mathbb{E}\left[\frac{b-X}{b-a}e^{\lambda a} + \frac{X-a}{b-a}e^{\lambda b} \mid \mathcal{F}\right] = \frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b}.$$

As in Lemma 14, define $\hat{\lambda} = \lambda(b-a)$, $p = \frac{-a}{b-a}$ and $f : \mathbb{R} \to \mathbb{R}$, $f(x) = -px + \log(1 - p + pe^x)$. Then

$$\frac{b}{b-a}e^{\lambda a} - \frac{a}{b-a}e^{\lambda b} = e^{f(\hat{\lambda})} \leq \exp\left(\frac{\lambda^2 (b-a)^2}{8}\right).$$

□

The following theorem is an extended version of Azuma's inequality. A similar theorem is proven in [6].

Theorem 15. *Let $X_0, X_1, X_2, \ldots, X_n$ be a martingale adapted to a filtration $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_n$, bounded by $a_i \leq X_i - X_{i-1} \leq b_i$ for all $i = 1, 2, \ldots, n$. Then for any $\epsilon > 0$*

$$\mathbb{P}(|X_n - X_0| \geq \epsilon) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right).$$

Proof. We start by proving the upper bound. By the exponential Markov inequality we obtain for all $\lambda > 0$

$$\mathbb{P}(X_n - X_0 \geq \epsilon) \leq e^{-\lambda\epsilon}\mathbb{E}\left[e^{\lambda(X_n - X_0)}\right].$$

In which

$$\begin{aligned}
&\mathbb{E}\left[e^{\lambda(X_n - X_0)}\right] \\
&\qquad = \mathbb{E}\left[e^{\lambda(X_n - X_{n-1})}e^{\lambda(X_{n-1} - X_0)}\right] \\
&\qquad = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda(X_n - X_{n-1})}e^{\lambda(X_{n-1} - X_0)} \mid \mathcal{F}_{n-1}\right]\right]
\end{aligned}$$

Observe that $e^{\lambda(X_{n-1} - X_0)}$ is $(\mathcal{F}_{n-1}, \mathcal{B})$-measurable. Therefore we can apply Theorem 9 to see that

$$\mathbb{E}\left[\mathbb{E}\left[e^{\lambda(X_n - X_{n-1})}e^{\lambda(X_{n-1} - X_0)} \mid \mathcal{F}_{n-1}\right]\right] = \mathbb{E}\left[\mathbb{E}\left[e^{\lambda(X_n - X_{n-1})} \mid \mathcal{F}_{n-1}\right]e^{\lambda(X_{n-1} - X_0)}\right]$$

With some basic properties and the definition of a martingale we obtain that

$$\begin{aligned}\mathbb{E}\left[X_n - X_{n-1} \mid \mathcal{F}_{n-1}\right] & \\ &= \mathbb{E}\left[X_n \mid \mathcal{F}_{n-1}\right] - \mathbb{E}\left[X_{n-1} \mid \mathcal{F}_{n-1}\right] \\ &= X_{n-1} - \mathbb{E}\left[X_{n-1} \mid \mathcal{F}_{n-1}\right] \\ &= X_{n-1} - X_{n-1} = 0.\end{aligned}$$

Also $a_n \leq X_n - X_{n-1} \leq b_n$. Therefore we can apply Corollary 14.1 to bound the expectation by

$$\mathbb{E}\left[\exp\left(\frac{1}{8}(\lambda(b_n - a_n))^2\right) e^{\lambda(X_{n-1} - X_0)}\right] = \exp\left(\frac{1}{8}(\lambda(b_n - a_n))^2\right) \mathbb{E}\left[e^{\lambda(X_{n-1} - X_0)}\right]$$

Using induction, we can expand this to

$$\exp\left(\frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right)$$

so that we obtain for our original equation:

$$\mathbb{P}\left(X_n - X_0 \geq \epsilon\right) \leq \exp\left(-\lambda\epsilon + \frac{\lambda^2}{8}\sum_{i=1}^{n}(b_i - a_i)^2\right).$$

To get the best bound, we would like to minimize the exponent for any ϵ. This is done by setting

$$\lambda = \frac{4\epsilon}{\sum_{i=1}^{n}(b_i - a_i)^2}$$

which proves the upper bound

$$\mathbb{P}\left(X_n - X_0 \geq \epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

The bound of $\mathbb{P}\left(X_0 - X_n \geq \epsilon\right)$ follows by symmetry, and the result follows by combining the lower and upper bound. □

The following definition plays a key role in the proof of McDiarmid's inequality.

Definition 25 (Doob martingale)**.** *Let $X_1, X_2, \ldots, X_n$ a stochastic process adapted to the filtration $\emptyset = \mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \subset \ldots \subset \mathcal{F}_n$, and let f a real valued function taking $X_1, X_2, \ldots, X_n$ as arguments. Then the Doob martingale of f with respect to the random variables $X_1, X_2, \ldots, X_n$ is defined as the sequence*

$$B_i = \mathbb{E}\left[f(X_1, X_2, \ldots, X_n) \mid \mathcal{F}_i\right]$$

for all $i = 0, 1, \ldots, n$.

Specifically the definition implies that $B_0 = \mathbb{E}[f(X_1, X_2, \ldots, X_n)]$ and that if $\mathcal{F}_n$ is the sigma-algebra included in the triple of the probability space, then $B_n = f(X_1, X_2, \ldots, X_n)$. For completeness, we provide the following theorem and proof.

Theorem 16. *The Doob martingale is a martingale with respect to its filtration.*

Proof. We will prove that the Doob martingale has the martingale property:

$$\begin{aligned}
&\mathbb{E}[B_{i+1} \mid \mathcal{F}_i] \\
&= \mathbb{E}[\mathbb{E}[f(X_1, X_2, \ldots, X_n) \mid \mathcal{F}_{i+1}] \mid \mathcal{F}_i] \\
&= \mathbb{E}[f(X_1, X_2, \ldots, X_n) \mid \mathcal{F}_i] \\
&= B_i
\end{aligned}$$

for all $i = 0, 1, \ldots, n-1$. □

The following theorem was first proven by McDiarmid in [6]. The proof here however differs from the proof in [6] as it applies the previously proven Azuma's extended inequality to a Doob martingale.

Theorem 17 (McDiarmid's inequality)**.** *Let $X = (X_1, X_2, \ldots, X_n)$ a vector of independent random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ a function such that replacing the i'th argument with an arbitrary value does not change the function more than $c_i \in \mathbb{R}$. That is,*

$$|f(x_1, \ldots, x_{i-1}, a_i, x_{i+1}, \ldots, x_n) - f(x_1, \ldots, x_{i-1}, b_i, x_{i+1}, \ldots, x_n)| \leq c_i$$

for all variables $x_1, x_2, \ldots, x_n$, a_i, b_i in all arguments $i = 1, 2, \ldots, n$.

Then

$$\mathbb{P}(|f(X) - \mathbb{E}[f(X)]| \geq \epsilon) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

Proof. Let $B_1, B_2, \ldots, B_n$ a Doob martingale of function f with respect to $X_1, X_2, \ldots, X_n$ and its natural filtration $\mathcal{F}_i$ for all $i = 1, 2, \ldots n$. We will show that this martingale has the assumptions required for Theorem 15. Therefore we are required to bound $B_i - B_{i-1}$ for all $i = 1, 2, \ldots, n$. It holds that

$$B_i - B_{i-1} = \mathbb{E}[f(X) \mid \mathcal{F}_i] - \mathbb{E}[f(X) \mid \mathcal{F}_{i-1}].$$

Let X_i' be an independent copy of X_i and let X_i' and X_1, X_2, …, X_n be independent. Define further $X' = (X_1, X_2, X_{i-1}, X_i', X_{i+1}, \ldots, X_n)$. It holds that X_i and X_i' are independent of $\mathcal{F}_{i-1}$ because $\mathcal{F}_{i-1}$ is the natural filtration of $X_1, X_2, \ldots, X_{i-1}$. Therefore by Theorem 11 it is true that

$$\mathbb{E}[f(X) \mid \mathcal{F}_{i-1}] = \mathbb{E}[f(X') \mid \mathcal{F}_{i-1}].$$

When plugging this into $B_i - B_{i-1}$ we obtain

$$\begin{aligned}
&B_i - B_{i-1} \\
&\qquad = \mathbb{E}[f(X) \mid \mathcal{F}_i] - \mathbb{E}[f(X') \mid \mathcal{F}_i] \\
&\qquad = \mathbb{E}[f(X) - f(X') \mid \mathcal{F}_i].
\end{aligned}$$

The constraint on the function f implies that there exists some $a_i, b_i \in \mathbb{R}$ such that $b_i - a_i = c_i$ and

$$a_i \leq f(X) - f(X') \leq b_i.$$

It follows that

$$a_i \leq \mathbb{E}\left[f(X) - f(X') \mid \mathcal{F}_i\right] = B_i - B_{i-1} \leq b_i$$

for all $i = 1, 2, 3, \ldots, n$. From Theorem 16 it follows that $B_1, B_2, B_3, \ldots, B_n$ is adapted to the filtration $(\mathcal{F}_i)_{i \in \{1,2,3,\ldots,n\}}$. Hence we can apply Theorem 15 to obtain

$$\mathbb{P}\left(|B_n - B_0| \geq \epsilon\right) = \mathbb{P}\left(|f(X) - \mathbb{E}\left[f(X)\right]| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2}{\sum_{i=1}^{n} c_i^2}\right).$$

□

An inequality theorem known as Hoeffding's inequality [4] follows directly from Mcdiarmid's inequality. This inequality provides concentration bounds for bounded independent random variables.

Theorem 18 (Hoeffding's inequality)**.** *Let $X_1, X_2, \ldots, X_n$ independent random variables such that $a_i \leq X_i \leq b_i$ for suitable constants $a_i, b_i \in \mathbb{R}$ for all $i = 1, 2, \ldots, n$. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. For all $\epsilon > 0$*

$$\mathbb{P}\left(|\overline{X} - \mathbb{E}\left[\overline{X}\right]| \geq \epsilon\right) \leq 2\exp\left(\frac{-2\epsilon^2 n^2}{\sum_{i=1}^{n}\left(b_i - a_i\right)^2}\right).$$

Proof. Observe that $\overline{X}$ is a function of $X_1, X_2, \ldots, X_n$ and is bounded in each coordinate $i = 1, 2, \ldots, n$ by $\frac{b_i - a_i}{n}$. □

In turn, Hoeffding's inequality is a generalization of a result first obtained by Chernoff [2]. This result provides concentration bounds for random variables with a Bernoulli distribution.

Theorem 19. *Let $X_1, X_2, \ldots, X_n$ independent and identical Bernoulli distributed random variables, where each variable takes on values in the set $\{0, 1\}$. Define $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$. For all $\epsilon > 0$*

$$\mathbb{P}\left(|\overline{X} - \mathbb{E}\left[\overline{X}\right]| \geq \epsilon\right) \leq 2e^{-2n\epsilon^2}.$$

Proof. Apply Hoeffding's inequality with $b_i = 1$, $a_i = 0$ for all $i = 1, 2, \ldots, n$. □

3.3 Chernoff bound

The following lemma and theorem are based on a paper by Chernoff[3].

Lemma 20 (Chernoff Lemma)**.** *Let $X_1, X_2, \ldots, X_N$ independent Bernoulli distributed. Let X the sum of $X_1, X_2, \ldots, X_N$. Then for all $t > 0$ and $a \geq 0$*

$$\mathbb{P}(X \geq a) \leq e^{-ta} \prod_i \mathbb{E}\left[e^{tX_i}\right].$$

Proof. By the extended Markov inequality we have for any $t > 0$ and $a \geq 0$

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}\left[e^{tX}\right]}{e^{ta}} = e^{-ta}\mathbb{E}\left[e^{tX}\right].$$

Because $X_1, X_2, \ldots, X_N$ are independent it holds that

$$\prod_i \mathbb{E}\left[e^{tX_i}\right] = \mathbb{E}\left[e^{tX}\right].$$

Therefore

$$\mathbb{P}(X \geq a) \leq e^{-ta} \prod_i \mathbb{E}\left[e^{tX_i}\right].$$

□

Theorem 21 (Multiplicative Chernoff bound)**.** *Let $X_1, X_2, \ldots, X_N$ independent Bernoulli distributed. Let $X = X_1 + X_2 + \ldots + X_N$. Then*

$$\mathbb{P}(X > (1+\delta)\mathbb{E}[X]) \leq \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}.$$

Proof. From Lemma 20 we immediately obtain for all $t > 0$

$$\mathbb{P}(X > (1+\delta)\mathbb{E}[X]) \leq \exp(t(1+\delta)\mathbb{E}[X]) \prod_i \mathbb{E}\left[e^{tX_i}\right].$$

Define $\mathbb{P}(X_i = 1) = p_i$ for all $i \in \{1, 2, \ldots, n\}$. This allows us to evaluate the expectation

$$\mathbb{E}\left[e^{tX_i}\right] = p_i e^t + (1 - p_i) = p_i(e^t - 1) + 1.$$

Observe that for all $x \geq 0$ it holds that $x + 1 \leq e^x$, which is evident when taking derivatives on both sides. When applying this result we see that

$$p_i(e^t - 1) + 1 \leq \exp(p_i(e^t - 1))$$

Hence we obtain for the product

$$\prod_i \mathbb{E}\left[e^{tX_i}\right] \leq \prod_i \exp(p_i(e^t - 1)) = \exp(\mathbb{E}[X](e^t - 1)).$$

We have arrived at

$$\begin{aligned}\mathbb{P}\left(X > (1+\delta)\mathbb{E}\left[X\right]\right) & \\ &\leq \exp(t(1+\delta)\mathbb{E}\left[X\right])\exp(\mathbb{E}\left[X\right](e^t-1)) \\ &= \exp(\mathbb{E}\left[X\right](t(1+\delta)+e^t-1)) \\ &= \exp((t(1+\delta)+e^t-1))^{\mathbb{E}[X]}.\end{aligned}$$

Setting $t = \log(1+\delta)$ gives

$$\exp((t(1+\delta)+e^t-1))^{\mathbb{E}[X]} = \left(\frac{e^\delta}{(1+\delta)^{1+\delta}}\right)^{\mathbb{E}[X]}.$$

This proves the desired result.

□

4 Statistical learning

Statistical learning (see [10] for an overview) is concerned with the finding of a function or algorithm with certain desired properties through learning on a training set. Potential candidate functions are named *classifiers*. In our set-up, we will look at independent and identically distributed samples with a distribution D over the domain $X \times Y$ where X denotes the set of input and Y the set of output which a desired classifier should have. Note that it is possible to observe multiple different output elements for the same input element. For example, assume that X is the number of wheels on a car and Y represents car colors. There exist multiple cars of different color with the same number of wheels. A classifier is a function $h : X \to Y$. We name the set of all classifiers H. To allow some granularity in determining whether or not a classifier is correct, we use a non-negative function $l : H \times X \times Y \to \mathbb{R}$, the *loss function*, which maps to each classifier for each input and output a value which indicates how far off a classifier is on that certain input, where 0 denotes no loss. For example

$$l(h,x,y) = \begin{cases} 0 & \text{if } h(x) = y \\ 1 & \text{otherwise} \end{cases}$$

which is called the 0-1 loss function.

4.1 PAC learnable

An important goal in statistical learning is to find a classifier $h \in H$ which has minimal loss. For this purpose we introduce a *statistical learning algorithm*: an algorithm which takes a finite number of samples and returns a classifier $h \in H$, determined entirely by those samples. We will use theory introduced by Valiant[8]. This theory aims to identify statistical learning algorithms which are able to find a classifier with minimal loss through sampling. This identifying

is done through a special statistical learning algorithm property: agnostic PAC learnable. The definition is given below. The definition requires that an agnostic PAC learnable algorithm finds a classifier with minimal loss except for two relaxations. The first one is that the algorithm is only approximately correct: the loss of the sample can be larger than the minimal loss, up to an *error* of ϵ. The second one is that there is a chance δ that nothing is guaranteed about the outcome of the algorithm. These limitations give rise to the name Probably Approximately Correct (PAC) [1].

Definition 26 (Agnostic PAC learnable)**.** *A statistical learning algorithm is called agnostic PAC learnable if for all $\epsilon, \delta > 0$ a $N \in \mathbb{N}$ exists such that for all distributions over $X \times Y$ we can run the algorithm on at least N i.i.d. samples returning a classifier $h \in H$ such that*

$$\mathbb{P}_{(x,y)\sim D}\left(\mathbb{E}\left[l(h,x,y)\right] \geq \min_{h' \in H} \mathbb{E}\left[l(h',x,y)\right] + \epsilon\right) \leq \delta.$$

Next, we will prove that any finite class H is agnostic PAC learnable through the Empirical Risk Minimizer (ERM) algorithm (see for example [9]).

First we introduce the sample average.

Definition 27 (Sample average)**.** *Let S a finite number of samples, and let h a classifier. Then define the sample average by*

$$\overline{l_S}(h) = \frac{1}{|S|} \sum_{(x,y)\in S} l(h,x,y).$$

Note that an algorithm can calculate the sample average in a finite amount of time if S is finite.

Definition 28 (ERM algorithm)**.** *The ERM algorithm is a statistical learning algorithm on finite H, which returns a classifier with minimal average sample error. That is, for a finite sample S it returns an h such that*

$$\overline{l_S}(h) \leq \overline{l_S}(h')$$

for all $h' \in H$.

The algorithm works by calculating for each $h \in H$ the average sample error, and then picking the h with the lowest average sample error.

Theorem 22. *If H is finite, and l is bounded, i.e. there exists $a, b \in \mathbb{R}$ such that $a \leq l(h,x,y) \leq b$ for all values of $h \in H, x \in X, y \in Y$, then the ERM algorithm is agnostic PAC learnable.*

Proof. Let us observe n i.i.d. samples $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\} = S \sim D^n$ for some $n \in \mathbb{N}$. Let $h \in H$ the returned function from executing the ERM algorithm on S. Then $l(h,x_1,y_2)$, $l(h,x_2,y_2)$, ..., $l(h,x_n,y_n)$ forms an i.i.d. sample. Additionally let $h' \in H - \{h\}$. Then it follows that $L = (L_i)_{i\in\{1,\ldots,n\}} = (l(h',x_i,y_i) - l(h,x_i,y_i))_{i\in\{1,\ldots,n\}}$ forms a set of i.i.d. random variables. However,

note that L is not bounded by a, b, but instead by $a - b \leq L_i \leq b - a$ for all $i = 1, 2, \ldots, n$. The sample average of L can be written by $\overline{L} = \overline{l_S}(h') - \overline{l_S}(h)$. We can apply Hoeffding's inequality to L

$$\mathbb{P}\left(|\overline{L} - \mathbb{E}\left[\overline{L}\right]| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2 n}{((a-b)-(b-a))^2}\right) = \exp\left(-\frac{\epsilon^2 n}{2(a-b)^2}\right)$$

It follows by symmetry that

$$\mathbb{P}\left(\overline{L} - \mathbb{E}\left[\overline{L}\right] \geq \epsilon\right) \leq \exp\left(-\frac{\epsilon^2 n}{2(b-a)^2}\right).$$

Note that

$$\begin{aligned}
&\mathbb{P}\left(\overline{L} - \mathbb{E}\left[\overline{L}\right] \geq \epsilon\right) \\
&\qquad = \mathbb{P}\left(\overline{l_S}(h') - \overline{l_S}(h) - \mathbb{E}\left[\overline{l_S}(h')\right] + \mathbb{E}\left[\overline{l_S}(h)\right] \geq \epsilon\right) \\
&\qquad = \mathbb{P}\left(\mathbb{E}\left[\overline{l_S}(h)\right] - \mathbb{E}\left[\overline{l_S}(h')\right] - \epsilon \geq \overline{l_S}(h) - \overline{l_S}(h')\right).
\end{aligned}$$

Because of the ERM algorithm, we have that $\overline{l_S}(h) - \overline{l_S}(h') \leq 0$. Therefore

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[\overline{l_S}(h)\right] - \mathbb{E}\left[\overline{l_S}(h')\right] - \epsilon \geq \overline{l_S}(h) - \overline{l_S}(h')\right) \\
&\qquad \leq \mathbb{P}\left(\mathbb{E}\left[\overline{l_S}(h)\right] - \mathbb{E}\left[\overline{l_S}(h')\right] - \epsilon \geq 0\right) \\
&\qquad \leq \mathbb{P}\left(\mathbb{E}\left[\overline{l_S}(h)\right] \geq \mathbb{E}\left[\overline{l_S}(h')\right] + \epsilon\right) \\
&\qquad \leq \exp\left(-\frac{\epsilon^2 n}{2(b-a)^2}\right).
\end{aligned}$$

Let $(x, y) \sim D$. Observe that $\mathbb{E}\left[\overline{l_S}(h)\right] = \mathbb{E}\left[l(h, x, y)\right]$ and similarly $\mathbb{E}\left[\overline{l_S}(h')\right] = \mathbb{E}\left[l(h', x, y)\right]$.

From this it follows that

$$\begin{aligned}
&\mathbb{P}\left(\mathbb{E}\left[l(h, x, y)\right] \geq \min_{h' \in H} \mathbb{E}\left[l(h', x, y)\right] + \epsilon\right) \\
&= \mathbb{P}\left(\text{there exists } h' \in H \text{ such that } \mathbb{E}\left[l(h, x, y)\right] \geq \mathbb{E}\left[l(h', x, y)\right] + \epsilon\right) \\
&\leq |H| \exp\left(-\frac{\epsilon^2 n}{2(b-a)^2}\right).
\end{aligned}$$

To finish the proof, define

$$\delta = |H| \exp\left(-\frac{\epsilon^2 n}{2(b-a)^2}\right)$$

it follows that when

$$n \geq \frac{2(b-a)^2}{\epsilon^2} \ln\left(\frac{|H|}{\delta}\right)$$

we have that

$$\mathbb{P}\left(\mathbb{E}\left[l(h, x, y)\right] \geq \min_{h' \in H} \mathbb{E}\left[l(h', x, y)\right] + \epsilon\right) \leq \delta$$

which is the definition of agnostic PAC learnable. □

4.2 Rademacher complexity

It is also possible to extend the concept of agnostic PAC learnable to certain infinite classes H. An elegant way to do this is by using the Rademacher Complexity.

Definition 29 (Rademacher complexity)**.** *Let H consist of real-valued classifiers and let $S = \{(x_1, y_1), (x_2, y_2), \dots (x_n, y_n)\}$ a sample with elements independently distributed by D and draw independently $\sigma_1, \sigma_2, \dots, \sigma_n$ from the Rademacher distribution, i.e. $\mathbb{P}(\sigma_i = 1) = \mathbb{P}(\sigma_i = -1) = \frac{1}{2}$. Then the empirical Rademacher complexity is defined by*

$$\hat{\mathcal{R}}_S(H) = \mathbb{E}\left[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \sigma_i l(h, x_i, y_i) \;\middle|\; S\right].$$

The Rademacher complexity is defined by

$$\mathcal{R}_S(H) = \mathbb{E}\left[\hat{\mathcal{R}}_S(H)\right] = \mathbb{E}\left[\sup_{h \in H} \frac{1}{n} \sum_{i=1}^{n} \sigma_i l(h, x_i, y_i)\right].$$

The following theorem shows the importance of the Rademacher complexity. It bounds the difference between the sample loss of a sample and the expectation of the sample loss by the sum of the Rademacher complexity and an arbitrary ϵ with probability going to 1 for n to infinity.

Theorem 23. *Let the loss function be bounded to $[0, 1]$ and let $S = \{(x_1, y_1), (x_2, y_2)\ , \dots\ , (x_n, y_n)\}$ a sample with elements independently distributed by D. Then for all $\epsilon \in (0, 1)$*

$$\mathbb{P}\left(\sup_{h \in H} \left\{\mathbb{E}\left[\overline{l_S}(h)\right] - \overline{l_S}(h)\right\} \geq 2\mathcal{R}_{\mathcal{S}}(\mathcal{H}) + \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2n}\right).$$

Proof. Pick some $h \in H$. Define $f(S) = \sup_{h \in H} \left\{\mathbb{E}\left[\overline{l_S}(h)\right] - \overline{l_S}(h)\right\}$. We will bound the difference of $f(S)$ and $\mathbb{E}[f(S)]$ using McDiarmid's inequality. Observe that $\mathbb{E}\left[\overline{l_S}(h)\right]$ does not depend on S and recall that $\overline{l_S}(h) = \frac{1}{n}\sum_{(x,y) \in S} l(h, x, y)$. Therefore, when replacing a single element in S by an arbitrary other element, $f(S)$ does not change more than $\frac{1}{n}$. This allows us to apply McDiarmid's inequality to f where we treat each element in S as a separate argument:

$$\mathbb{P}\left(|f(S) - \mathbb{E}[f(S)]| \geq \epsilon\right) \leq 2\exp\left(\frac{-\epsilon^2}{2n}\right).$$

It follows that

$$\mathbb{P}\left(f(S) \geq \mathbb{E}[f(S)] + \epsilon\right) \leq \exp\left(\frac{-\epsilon^2}{2n}\right).$$

Hence to prove the theorem it suffices to show that $\mathbb{E}[f(S)] \leq 2\mathcal{R}_{\mathcal{S}}(\mathcal{H})$. Let $S' = \{(x'_1, y'_1), (x'_2, y'_2), \ldots, , (x'_n, y'_n)\}$ be an independent copy of S. Then

$$\begin{aligned} & f(S) \\ & \quad = \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_S}(h)\right] - \overline{l_S}(h) \right\} \\ & \quad = \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h)\right] - \overline{l_S}(h) \right\} \\ & \quad = \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h)\right] - \mathbb{E}\left[\overline{l_S}(h) \mid S\right] \right\} \end{aligned}$$

Because S' is an independent copy of S, we have that S and $\overline{l_{S'}}(h)$ are independent for all $h \in H$. So by Theorem 10 (pulling out independent factors) we see that

$$\begin{aligned} & \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h)\right] - \mathbb{E}\left[\overline{l_S}(h) \mid S\right] \right\} \\ & \qquad = \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h) \mid S\right] - \mathbb{E}\left[\overline{l_S}(h) \mid S\right] \right\} \\ & \qquad = \sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h) - \overline{l_S}(h) \mid S\right] \right\}. \end{aligned}$$

For all $h \in H$ it holds that

$$\mathbb{E}\left[\overline{l_{S'}}(h) - \overline{l_S}(h) \mid S\right] \leq \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_{S'}}(h) - \overline{l_S}(h) \right\} \mid S \right].$$

Therefore

$$\sup_{h \in H} \left\{ \mathbb{E}\left[\overline{l_{S'}}(h) - \overline{l_S}(h) \mid S\right] \right\} \leq \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_{S'}}(h) - \overline{l_S}(h) \right\} \mid S \right].$$

We have established that

$$f(S) \leq \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_{S'}}(h) - \overline{l_S}(h) \right\} \mid S \right].$$

Therefore we have in expectation that

$$\mathbb{E}[f(S)] \leq \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_{S'}}(h) - \overline{l_S}(h) \right\} \right].$$

Introduce the independent random variables $\sigma_1, \sigma_2, \ldots, \sigma_n$ distributed by the Rademacher distribution. Then

$$\begin{aligned} & \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_{S'}}(h) - \overline{l_S}(h) \right\} \right] \\ & \qquad = \mathbb{E}\left[\sup_{h \in H} \left\{ \overline{l_S}(h) - \overline{l_{S'}}(h) \right\} \right] \\ & \qquad = \mathbb{E}\left[\sup_{h \in H} \left\{ \frac{1}{n} \sum_{i=1}^{n} l(h, x_i, y_i) - l(h, x'_i, y'_i) \right\} \right] \end{aligned}$$

Define the samples

$$u_i = \begin{cases} (x_i, y_i) & \text{if } \sigma_i = 1 \\ (x'_i, y'_i) & \text{otherwise} \end{cases}$$

and

$$u'_i = \begin{cases} (x'_i, y'_i) & \text{if } \sigma_i = 1 \\ (x_i, y_i) & \text{otherwise} \end{cases}$$

for all $i = 1, 2, \ldots, n$. Define $U = (u_1, u_2, \ldots, u_n)$ and $U' = (u'_1, u'_2, \ldots, u'_n)$. Note that U and U' together contain all the elements contained in S and S' together. As stated, (x_i, y_i) and (x'_i, y'_i) are identically and independently distributed for all $i = 1, 2, \ldots, n$. Therefore we can swap the sample sets S and S' by U and U' and still obtain the same result in expectation. More formally:

$$\mathbb{E}\left[\sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} l(h, x_i, y_i) - l(h, x'_i, y'_i)\right\}\right] = \mathbb{E}\left[\sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n}(l(h, u_i) - l(h, u'_i))\right\}\right].$$

This is equal to

$$\mathbb{E}\left[\sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(l(h, x_i, y_i) - l(h, x'_i, y'_i))\right\}\right].$$

It follows that

$$\begin{aligned} &\mathbb{E}\left[f(S)\right] \\ &\quad \leq \mathbb{E}\left[\sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} \sigma_i(l(h, x_i, y_i) - l(h, x'_i, y'_i))\right\}\right] \\ &\quad \leq \mathbb{E}\left[\sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} \sigma_i l(h, x_i, y_i)\right\} + \sup_{h \in H}\left\{\frac{1}{n}\sum_{i=1}^{n} -\sigma_i l(h, x_i, y_i)\right\}\right]. \end{aligned}$$

Observe that $-\sigma_i$ is also distributed by the Rademacher distribution for all $i = 1, 2, \ldots, n$. Therefore we have by the definition of the Rademacher complexity:

$$\mathbb{E}\left[f(S)\right] \leq 2\mathcal{R}_{\mathcal{S}}(\mathcal{H}).$$

□

5 Random geometric traveling salesman problem

The traveling salesman problem can be stated as follows. Define the points $P_1, P_2, \ldots, P_n \in \mathbb{R}^2$ and the set $P = \{P_1, P_2, \ldots, P_n\}$. Give an order, known as a *tour* $T = (t_1, t_2, \ldots, t_n)$ in which to visit all points in P such that the total of all distances between each two sequential points is minimized, i.e.

$\sum_{i=2}^{n} d(P_{t_{i-1}}, P_{t_i})$ is minimal, where d is the euclidean metric. In the random geometric version, the points are independent random variables, uniformly distributed over the unit square $[0,1]^2$. The length of the shortest tour can be expressed as $t : ([0,1]^2)^n \to \mathbb{R}$. We will show that the length $t(P)$ of the shortest tour is concentrated strongly around its mean.

We start with a result from McDiarmid [6].

Theorem 24. *For all $\epsilon > 0$, it holds that*

$$\mathbb{E}\left[|t(P) - \mathbb{E}\left[t(P)\right]| \geq \epsilon\sqrt{n}\right) \leq 2\exp\left(\frac{-\epsilon^2}{4}\right)\right]$$

Proof. We will prove this theorem by applying McDiarmid's inequality. Let $\delta > 0$. Observe that the distance between two points lying in the unit square can never be larger than $\sqrt{2}$. Hence the difference in t arising from a move of a single point can never be larger than $2\sqrt{2}$. Therefore McDiarmid's inequality gives

$$\mathbb{P}\left(|t(P) - \mathbb{E}\left[t(P)\right]| \geq \delta\right) \leq 2\exp\left(\frac{-2\delta^2}{\sum_{i=1}^{n}(2\sqrt{2})^2}\right) = 2\exp\left(\frac{-\delta^2}{4n}\right).$$

Substitute $\delta = \epsilon\sqrt{n}$ for the desired result. □

An improved bound has been discovered by Rhee and Talagrand [7]. This bound uses Azuma's inequality. First we prove a useful lemma.

Lemma 25. *Let $P = \{P_1, P_2, \ldots, P_n\}$ a set of independent random variables, uniformly distributed over the unit square. For some point $x \in [0,1]^2$, let $D(P,x)$ be the distance between x and the closest point in P. Then*

$$\mathbb{E}\left[D(P,x)\right] \leq \frac{2}{\sqrt{n}}.$$

Proof. Let $x \in [0,1]^2$. Denote a ball centered at x with radius $r \in [0,\sqrt{2}]$ by $B(x,r)$. Let A be the area of $B(x,r) \cap [0,1]^2$. We have that $D(P,x) > r$ precisely when all of the points $P_1, P_2, \ldots, P_n$ are outside of the area A. Therefore

$$\mathbb{P}\left(D(P,x) > r\right) = (1-A)^n.$$

We will determine an expression of the form $A \geq cr^2$ for some constant $c \in \mathbb{R}$, which holds for all x and r. The area A can be decreased by moving x to one of the four corners of the unit square. Therefore assume without generality that x is at one of the four corners of the unit square. By symmetry all corners are equivalent. We will handle three cases for r, which are illustrated in Figure 1. Assume $r \leq 1$. In this case precisely a quarter of the circle $B(x,r)$ intersects with the unit square, and as such $A = \frac{\pi}{4}r^2$. Assume $1 < r < \sqrt{2}$. We have that of the quarter circle a part falls out of the unit square. Therefore $A < \frac{\pi}{4}r^2$, and without loss of generality we can state that $r > 1$. As r increases, the overlap

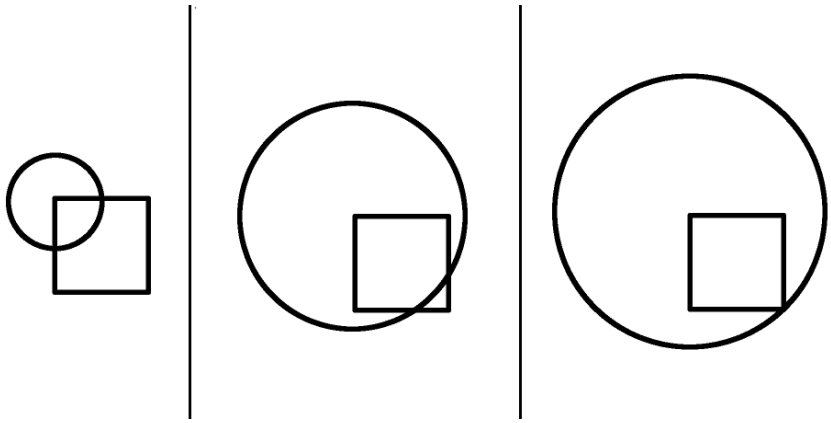

Figure 1: The three cases of $B(x, r)$ intersecting with the unit square. On the left, $r \leq 1$. In the middle, $1 < r < \sqrt{2}$. On the right, $r = \sqrt{2}$.

relative to r^2 decreases. Therefore we only need to consider $r = \sqrt{2}$. If $r = \sqrt{2}$, the unit square falls entirely inside $B(x, r)$, and we have that $A = 1 = \frac{1}{2}r^2$. Therefore $c = \frac{1}{2}$.

It follows that $A \geq \frac{1}{2}r^2 = cr^2$. So,

$$\mathbb{P}\left(D(P, x) > r\right) \leq (1 - cr^2)^n.$$

We will apply Fubini's theorem, which states that under certain conditions we can can change the order of integration. Let $f : \mathbb{R} \to \mathbb{R}$ the probability density function of $D(P, x)$. Then

$$\begin{aligned}
&\int_{[0,\infty)} \mathbb{P}\left(D(P, x) > r\right)\, \mathrm{d}t \\
&\qquad = \int_{[0,\infty)} \int_{[t,\infty)} f(s)\, \mathrm{d}s\, \mathrm{d}t \\
&\qquad = \int_{[0,\infty)} \int_{[0,\infty)} 1_{[t,\infty)}(s) f(s)\, \mathrm{d}s\, \mathrm{d}t \\
&\qquad = \int_{[0,\infty)} \int_{[0,\infty)} 1_{[t,\infty)}(s) f(s)\, \mathrm{d}t\, \mathrm{d}s \\
&\qquad = \int_{[0,\infty)} f(s) \int_{[0,\infty)} 1_{[t,\infty)}(s)\, \mathrm{d}t\, \mathrm{d}s \\
&\qquad = \int_{[0,\infty)} f(s) \int_{[0,\infty)} \begin{cases} 1 & \text{if } s \geq t \\ 0 & \text{otherwise} \end{cases} \mathrm{d}t\, \mathrm{d}s \\
&\qquad = \int_{[0,\infty)} s f(s)\, \mathrm{d}s \\
&\qquad = E[D(P, x)].
\end{aligned}$$

So we have that

$$\mathbb{E}\left[D(P,x)\right] \leq \int_0^{\sqrt{2}} (1-cr^2)^n \,\mathrm{d}r.$$

Because $1 - cr^2 \leq e^{-cr^2}$ we have that

$$\int_0^{\sqrt{2}} (1-cr^2)^n \,\mathrm{d}r. \leq \int_0^{\sqrt{2}} e^{-cnr^2} \,\mathrm{d}r.$$

With substituting we see that

$$\int_0^{\sqrt{2}} e^{-cnr^2} \,\mathrm{d}r = \int_0^{\sqrt{2}} e^{-(\sqrt{cn}r)^2} \,\mathrm{d}r = \int_0^{\sqrt{\frac{2}{cn}}} e^{-r^2} \,\mathrm{d}r \leq \sqrt{\frac{2}{cn}}.$$

With the identity $c = \frac{1}{2}$ we see that

$$\mathbb{E}\left[D(P,x)\right] \leq \frac{2}{\sqrt{n}}.$$

□

Theorem 26. *Define the constant* $c = \exp(\frac{1}{8\sqrt{2}} + 1)$. *Then for all* $\epsilon > 0$

$$\mathbb{P}\left(|t(P) - \mathbb{E}\left[t(P)\right]| \geq \epsilon\sqrt{n}\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{8\ln(c(n-1))}\right).$$

Proof. Define for each $i \in \{1,2,\ldots,n\}$ $T_i = \mathbb{E}\left[t(P) \mid P_1, P_2, \ldots, P_i\right]$. Further define the set $P' = \{P_1, \ldots, P_{i-1}, P_i', P_{i+1}, \ldots, P_n\}$, where P_i' is an independent copy of P_i and independent of all other $P_1, P_2, \ldots, P_n$, i.e. an independent uniform random point in the unit square. Then

$$\begin{aligned} &T_i - T_{i-1} \\ &\qquad = \mathbb{E}\left[t(P) \mid P_1, P_2, \ldots, P_i\right] - \mathbb{E}\left[t(P') \mid P_1, P_2, \ldots, P_{i-1}\right] \\ &\qquad = \mathbb{E}\left[t(P) \mid P_1, P_2, \ldots, P_i\right] - \mathbb{E}\left[t(P') \mid P_1, P_2, \ldots, P_i\right] \\ &\qquad = \mathbb{E}\left[t(P) - t(P') \mid P_1, P_2, \ldots, P_i\right]. \end{aligned}$$

It follows that

$$|T_i - T_{i-1}| \leq \mathbb{E}\left[\ |t(P) - t(P')|\ \mid P_1, P_2, \ldots, P_i\right].$$

Let $i \in \{1,2,\ldots,n-1\}$. Define $P^- = P - \{P_i\} = P' - \{P_i'\}$. Consider the increase in length for the shortest tour through P^- when adding a point $x \in [0,1]^2$. This point can be added to the tour by adding at most two edges between x and the closest point in P^-. The length of these edges is given by $D(P^-, x)$. Therefore

$$t(P^-) \leq t(P) \leq t(P^-) + 2D(P^-, P_i)$$

and

$$t(P^-) \leq t(P') \leq t(P^-) + 2D(P^-, P_i').$$

By combining these inequalities we obtain

$$-2D(P^-, P_i') \leq t(P) - t(P') \leq 2D(P^-, P_i).$$

So

$$-\mathbb{E}\left[2D(P^-, P_i') \mid P_1, P_2, \ldots, P_i\right] \leq \mathbb{E}\left[t(P) - t(P') \mid P_1, P_2, \ldots, P_i\right]$$

$$\leq \mathbb{E}\left[2D(P^-, P_i) \mid P_1, P_2, \ldots, P_i\right].$$

Observe that by Theorem 11, in the expectation

$$\mathbb{E}\left[2D(P^-, P_i') \mid P_1, P_2, \ldots, P_i\right]$$

all variables $P_i', P_{i+1}, P_{i+2}, \ldots, P_n$ can be replaced by independent copies. So we have that

$$\begin{aligned}\mathbb{E}\left[2D(P^-, P_i') \mid P_1, P_2, \ldots, P_i\right] & \\ &\leq \mathbb{E}\left[2D(\{P_{i+1}, P_{i+2}, \ldots, P_n\}, P_i') \mid P_1, P_2, \ldots, P_i\right] \\ &= \mathbb{E}\left[2D(\{P_{i+1}, P_{i+2}, \ldots, P_n\}, P_i')\right].\end{aligned}$$

Indeed we can apply Lemma 25 for the points $P_{i+1}, P_{i+2}, \ldots, P_n$ and we obtain that

$$-\frac{4}{\sqrt{n-i}} \leq \mathbb{E}\left[t(P) - t(P') \mid P_1, P_2, \ldots, P_i\right] \leq \frac{4}{\sqrt{n-i}}$$

and so for all $i < n$

$$|T_i - T_{i-1}| \leq \frac{4}{\sqrt{n-i}}.$$

From Theorem 24 we obtain that for all $i \in \{1, 2, \ldots, n\}$, specifically $i = n$, it holds that

$$|T_i - T_{i-1}| \leq 2\sqrt{2}.$$

Finally, we can apply Theorem 15 to find that for all $\epsilon > 0$

$$\begin{aligned}\mathbb{P}\left(\left|t(P) - \mathbb{E}\left[t(P)\right]\right| \geq \epsilon\right) & \\ &= \mathbb{P}\left(|T_n - T_0| \geq \epsilon\right) \\ &\leq 2\exp\left(\frac{-2\epsilon^2}{2\sqrt{2} + \sum_{i=1}^{n-1} \frac{16}{n-i}}\right) \\ &= 2\exp\left(\frac{-\epsilon^2}{\sqrt{2} + \sum_{i=1}^{n-1} \frac{8}{i}}\right) \\ &\leq 2\exp\left(\frac{-\epsilon^2}{\sqrt{2} + 8 + 8\ln(n-1)}\right).\end{aligned}$$

Set $c = \exp(\frac{1}{8\sqrt{2}} + 1)$ to obtain

$$\mathbb{P}\left(|t(P) - \mathbb{E}\left[t(P)\right]| \geq \epsilon\sqrt{n}\right) \leq 2\exp\left(\frac{-\epsilon^2 n}{8\ln(c(n-1))}\right).$$

□

References

[1] Dana Angluin. Queries and concept learning. *Mach. Learn.*, 2(4):319–342, April 1988.

[2] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statistics*, 23:493–507, 1952.

[3] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.*, 23(4):493–507, 12 1952.

[4] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, 58:13–30, 1963.

[5] W. A. J. Luxemburg. Arzelà's dominated convergence theorem for the Riemann integral. *Amer. Math. Monthly*, 78:970–979, 1971.

[6] Colin McDiarmid. On the method of bounded differences. In *Surveys in combinatorics, 1989 (Norwich, 1989)*, volume 141 of *London Math. Soc. Lecture Note Ser.*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[7] WanSoo T. Rhee and Michel Talagrand. Martingale inequalities and NP-complete problems. *Math. Oper. Res.*, 12(1):177–181, 1987.

[8] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.

[9] V. N. Vapnik. Principles of Risk Minimization for Learning Theory. *Advances in Neural Information Processing Systems*, 4:831–838, 1992.

[10] U. von Luxburg and B. Schölkopf. *Statistical Learning Theory: Models, Concepts, and Results*, volume 10, pages 651–706. Elsevier North Holland, Amsterdam, Netherlands, May 2011.