AZUMA'S INEQUALITY

WILLIAM CLARK

1. Introduction

The goal of this article is to give the reader exposure to the measure theory that is necessary to rigorously state and prove Azuma's Inequality, and show the technique that is used to prove the inequality. Here is what we will work up to:

Azuma's Inequality: If $X = (M_n)_{n\in\mathbb{N}_0}$ is a martingale with respect to the filtration $\mathbb{F} = \sigma(X)$ with $M_0 = 0$ and if there is a sequence $(c_k)_{k\in\mathbb{N}}$ of nonnegative numbers with $|M_n - M_{n-1}| \leq c_n$ for all $n \in \mathbb{N}$, then for all $\lambda \geq 0$ we have

$$\mathbb{P}[|M_n| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_{k=1}^{n} {c_k}^2}\right).$$

In plain language, Azuma's inequality gives an exponential bound on the probability that the nth random variable in a sequence of random variables takes on a value outside of a centered interval, given that the sequence satisfies certain properties. The main property that the sequence should satisfy is the martingale property. Informally, this property means that the predicted value of mth random variable when given information about the prior $m-1$ random variables is the same as the predicted value of the $m-1$th random variable. One can imagine how this might apply to the stock market, where the expected movement of today's stock is the same as the expected movement of yesterday's stock, despite the movement that occured on the stock market yesterday.

2. Measure and Probability Space

In this section we give the axioms of probability theory. These will all be necessary to rigorously describe random variables and martingales.

Assume $\Omega \neq \emptyset$ is an arbitrary nonempty set, and that 2^Ω is the set of all subsets (known as the power set) of Ω.

Definition 2.1 (σ-algebra)**.** *Let Ω be a set. Then a σ-algebra $\mathcal{F}$ is a non-empty collection of subsets of Ω such that the following hold:*

(1) $\Omega \in \mathcal{F}$.

(2) If $A \in \mathcal{F}$, then so is the complement of A, denoted A^c. (We have $A^c = \Omega - A$).

(3) If A_n is a sequence of elements of $\mathcal{F}$, then

$$\bigcup_{i=1}^{\infty} A_n \in \mathcal{F}.$$

(Definition 1.2 in [2]*)*

Date: December 7, 2021.

As we will see, σ-algebras are ubiquitous throughout the theory of probability. This warrants an example:

Example 1. Let $\Omega = \{1, 2, 3\}$. Let us verify that $\mathcal{F} = \{\emptyset, \{1, 2, 3\}\}$ is a σ-algebra, and that $\mathcal{A} = \{\emptyset, \{1, 2, 3\}, \{1\}, \{2\}, \{3\}\}$, is not a σ-algebra.

1. $\Omega = \{1, 2, 3\} \in \mathcal{F}$, as desired. Also, $\Omega \in \mathcal{A}$.
2. $\Omega^c = \emptyset \in \mathcal{F}$ and $\emptyset^c = \Omega \in \mathcal{F}$, as desired. However, $\{1\}^c = \{2, 3\} \notin \mathcal{A}$, which shows that $\mathcal{A}$ is not a σ-algebra.
3. Any sequence of elements chosen from $\mathcal{F}$, such as $\emptyset, \emptyset, \Omega, \cdots$ will only have $\emptyset$ and Ω, and $\emptyset \cup \Omega = \Omega \in \mathcal{F}$, as desired.

Definition 2.2 (probability measure, measure)**.** *Let $\Omega \neq \emptyset$ be an arbitrary set. Let $\mathcal{A} \subset 2^\Omega$ and let $\mu : \mathcal{A} \to [0, \infty]$ be a function which sends each set in $\mathcal{A}$ to a number in the interval $[0, \infty]$. We say that μ is a probability measure if we have the following conditions:*

(i) $\mu(\Omega) = 1$.

(ii) *$\mathcal{A}$ is a σ-algebra.*

(iii) *μ is σ-additive. That is, if $A_1, A_2, \cdots \in \mathcal{A}$ is a choice of countably many mutually disjoint sets in $\mathcal{A}$, then*

$$\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i).$$

(Note how (ii) implies that $\bigcup_{i=1}^{\infty} A_i \in A$.)

We say that μ is a measure if μ satisfies conditions (ii) and (iii). (Definition 1.28 in [2]*)*

Notice the striking similarities between the above definition, and the axioms of probability that were given in MAT340. Clearly, Ω acts as the sample space, and μ acts as the probability function p, assigning each sample point a probability.

Definition 2.3 (measurable space, measurable sets, discrete)**.** *A pair $(\Omega, \mathcal{A})$ consisting of a nonempty set Ω and a σ-algebra $\mathcal{A} \subset 2^\Omega$ is called a measurable space. The sets $A \in \mathcal{A}$ are called measurable sets. If Ω is at most countably infinite and if $\mathcal{A} = 2^\Omega$, then the measurable space $(\Omega, 2^\Omega)$ is called discrete. (Definition 1.38 (i) in* [2]*)*

Consider a coin flip in the context of the above definition:

Example 2. Let $\Omega = \{H, T\}$. Let the σ-algebra $\mathcal{A} = \{\emptyset, \{H, T\}, \{T\}, \{H\}\}$. Then $(\Omega, \mathcal{A})$ is a measurable space by the above definition. Further, $\{H\}$ is a measurable set, along with $\{T\}, \{T, H\}, \emptyset$. Further, since $\mathcal{A} = 2^{\{H,T\}}$, we call $(\Omega, \mathcal{A})$ discrete. If $\mathcal{F} = \{\{H\}, \{H, T\}, \emptyset\}$, consider the intuition behind why $(\Omega, \mathcal{F})$ is *not* a measurable space. Well, $\{T\} \notin \mathcal{F}$, so we cannot assign a probability to the event that $\{H\}$ does not happen, namely $\mu(\{H\}^c)$ is undefined. Thus, the notion of a measurable space agrees with the way that we combine and consider complements of probabilistic events.

Definition 2.4 (measure space)**.** *A triple $(\Omega, \mathcal{A}, \mu)$ is called a measure space if $(\Omega, \mathcal{A})$ is a measurable space and if μ is a measure on $\mathcal{A}$. (Definition 1.38 (ii) in* [2]*)*

We arrive at the formal definition of a probability space.

Definition 2.5 (probability space)**.** *A measure space $(\Omega, \mathcal{A}, \mathbb{P})$ such that $\mathbb{P}$ is a probability measure is called a probability space. The sets $A \in \mathcal{A}$ are called events. (Definition 1.38 (iii) in* [2]*)*

Having defined probability space, it will be easier to concretely state the other notions from probability that will be necessary for Azuma's inequality, and we will eliminate ambiguities surrounding the question of which measure, sample space, and event combinations are in consideration.

3. Maps and Random Variables

For the enthusiastic math student, learning that probability theory utilizes topology will come as an exciting surprise. It turns out that random variables, which in MAT340 we understood as maps from a sample space to $\mathbb{R}$, can be stated as maps from elements in one measurable space to elements in another measurable space. The codomain that will be most common will be the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, which is essentially the codomain the random variables in MAT340 had. We need topology to define $\mathcal{B}(\mathbb{R})$.

Definition 3.1 (topology, topological space)**.** *Let $\Omega \neq \emptyset$ be an arbitrary set. A class of sets $\tau \subset 2^{\Omega}$ is called a topology on Ω is it has the following three properties:*

(i) $\emptyset, \Omega \in \tau$.
(ii) $A \cap B \in \tau$ for any $A, B \in \tau$.
(iii) $(\bigcup_{A \in \mathcal{F}} A) \in \tau$ for any $\mathcal{F} \subset \tau$.

The pair (Ω, τ) is called a topological space. The sets $A \in \tau$ are called open, and the sets $A \subset \Omega$ with $A^c \in \tau$ are called closed. (Definition 1.20 in [2]*)*

We will leave it to the reader to determine the similarities and differences between a σ-algebra and a topology. For the next definition, we will be most focused on the case of $\Omega = \mathbb{R}$.

In the following definition, note that

$$\sigma(\tau) = \bigcap_{\substack{\mathcal{A} \subset 2^{\Omega} \text{ is a } \sigma\text{-algebra} \\ \mathcal{A} \subset \tau}} \mathcal{A}.$$

See Theorem 1.16 in [2] for more.

Definition 3.2 (Borel σ-algebra)**.** *Let (Ω, τ) be a topological space. The σ-algebra*

$$\mathcal{B}(\Omega) := \mathcal{B}(\Omega, \tau) := \sigma(\tau)$$

that is generated by the open sets is called the Borel σ-algebra on Ω. The elements $A \in \mathcal{B}(\Omega, \tau)$ are called Borel Sets or Borel measurable sets. (Definition 1.21 in [2]*)*

As is probably apparent, it is hard to grasp the arbitrary unions of uncountable sets. Thus, we will note that there are countable bases of Borel σ-algebras described in [2] that allow these sets to be more manageable. However, our goal is to lay down the definitions necessary, and so we will move forward. For convenience we include the definition of inverse image, which is not to be confused with inverse function.

Definition 3.3 (Inverse Image)**.** *Let $f : X \to Y$ be a function from a set X to a set Y. If U is a subset of Y, we define the set $f^{-1}(U)$ to be the set*

$$f^{-1}(U) := \{x \in X \mid f(x) \in U\},$$

and call this the inverse image of U. Note that it is not necessary for the inverse function of f to exist to make sense of the inverse image of U. (Definition 3.4.5 in [4]*)*

Take account of the use of the inverse image in the definition for a measurable map:

Definition 3.4 (measurable maps)**.** *Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces.*

(i) A map $X : \Omega \to \Omega'$ is called $\mathcal{A}-\mathcal{A}'$-measurable (or, briefly, measurable) if $X^{-1}(\mathcal{A}') := \{X^{-1}(A') : A' \in \mathcal{A}'\} \subset \mathcal{A}$; that is, if

$$X^{-1}(A') \in \mathcal{A} \text{ for any } A' \in \mathcal{A}'.$$

If X is measurable, we write $X : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$.

(ii) If $\Omega' = \mathbb{R}$ and $\mathcal{A}' = \mathcal{B}(\mathbb{R})$ is the Borel σ-algebra on $\mathbb{R}$, then $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is called an $\mathcal{A}-$measurable real map. (Definition 1.76 in [2]*)*

Now we have defined both the notion of a measurable set and a measurable map. The motivation for a measurable set was clear: we wanted it to fit the ways we combine and make complements of probabilistic events. Consider how the definition of a measurable map agrees with what we want a random variable to be.

Definition 3.5 (random variables)**.** *Let $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ be measurable spaces and let $X : \Omega \to \Omega'$ be measurable.*

(i) X is called a random variable with values in $(\Omega', \mathcal{A}')$. If $(\Omega', \mathcal{A}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then X is called a real random variable or simply a random variable. (Definition 1.102 in [2]*)*

Example 1. Let $X : (\{H, T\}, \mathcal{A} = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}) \to (\{0, 1\}, \mathcal{F} = \{\{0\}, \{1\}, \{0, 1\}, \emptyset\})$ be the random variable that counts the number of heads for one coin flip. Thus $X(H) = 0$ and $X(T) = 1$. We want X to be measurable, in particular because X is a very reasonable example of a random variable. Note first that $\mathcal{A}$ and $\mathcal{F}$ are σ-algebras, which means that the domain and codomain are each measurable spaces. Observe:

$$X^{-1}(\{0\}) = \{T\} \in \mathcal{A}$$
$$X^{-1}(\{0, 1\}) = \{H, T\} \in \mathcal{A}$$
$$X^{-1}(\emptyset) = \emptyset \in \mathcal{A}$$
$$X^{-1}(\{1\}) = \{H\} \in \mathcal{A},$$

demonstrating that X is measurable, as expected.

4. Expectation

The notion of a martingale cannot be expressed without the notion of the expected value of a random variable. As we will see, the definition of expectation relies on the integral. Measure theory gives us a general notion of an integral that works for both discrete and continuous cases. We now walk through the construction of the integral.

For this section let $(\Omega, \mathcal{A}, \mu)$ be a measure space. Further, given an arbitrary set A and an arbitrary set B, the indicator function $\mathbb{1}_A : B \to \{0, 1\}$ is defined by $\mathbb{1}_A(b) = 1$ if $b \in A$ and $\mathbb{1}_A(b) = 0$ if $b \notin A$, for all $b \in B$. Also, given two functions f and g with domain set A, we write $f = g$ given that $f(a) = g(a)$ for all $a \in A$. Also, we have that the sum of two functions is defined by $(g + f)(a) = g(a) + f(a)$ for all $a \in A$. This is extended to numerous functions in a similar manner.

Definition 4.1 (simple functions)**.** *Let $(\Omega, \mathcal{A})$ be a measurable space. A map $f : \Omega \to \mathbb{R}$ is called a simple function if there is an $n \in \mathbb{N}$ and mutually disjoint measurable sets $A_1, \cdots, A_n \in \mathcal{A}$, as well as numbers $\alpha_1, \cdots, \alpha_n \in \mathbb{R}$, such that*

$$f = \sum_{i=1}^{n} \alpha_i \mathbb{1}_{A_i}.$$

(Definition 1.93 in [2]*)*

Since the simple function plays a monumental role in the construction of the integral, it is worth reflecting on its definition. In particular, the image of a simple function is a set of cardinality at most $n+1$. It is remarkable that functions with finite images play a role in the integral, which we normally associate with continuous objects.

As a reminder for the next definition, a vector space is a set of vectors together with a field ($\mathbb{R}$ is usually the field in use) that is closed under scalar multiplication (the scalars are the elements of the field), closed under addition, and has a few additional properties. However, we will not need the full definition of a vector space to move forward.

Notation 4.1 $(\mathbb{E}, \mathbb{E}^+)$**.** *We let $\mathbb{E}$ be the vector space of simple functions on $(\Omega, \mathcal{A})$. Further, we let*

$$\mathbb{E}^+ := \{f \in \mathbb{E} : f \geq 0\}$$

be the set (not vector space) of nonnegative simple functions on $(\Omega, \mathcal{A})$.

Note that $f \geq 0$ means that if $x \in \operatorname{dom}(f)$, then $f(x) \geq 0$. Hence, non-negative simple functions will have nonnegative scalars. Further, recall that we are working in the measure space $(\Omega, \mathcal{A}, \mu)$, such that μ is a measure.

Definition 4.2 (normal representation, the map I)**.** *If the summation from definition 4.1 has scalars $\alpha_1, \cdots, \alpha_n \in (0, \infty)$, then that summation is called the normal representation of f. Next, we define the map $I : \mathbb{E}^+ \to [0, \infty]$ by*

$$I(f) = \sum_{i=1}^{n} \alpha_i \mu(A_i)$$

if f has the normal representation suggested above. (Definition 4.2 in [2]*)*

Thus, the map I takes simple functions to numbers by making use of the measure μ. Note that the normal representation of a simple function is unique, thus each simple function is mapped to exactly one value [2].

Recall that sup denotes the least upper bound of a set of real numbers. The following is the surprising definition of the integral for functions with positive codomains. Take account of the vastly different presentation in comparison to standard calculus texts.

Definition 4.3 (integral)**.** *If $f : \Omega \to [0, \infty]$ is measurable, then we define the integral of f with respect to μ by*

$$\int f d\mu := \sup\{I(g) \mid g \in \mathbb{E}^+, g \leq f\}.$$

(Definition 4.4 in [2]*)*

This definition says that the integral of f with respect to the measure μ is the least upper bound of the image $I(G)$, where $G = \{g \in \mathbb{E}^+, g \leq f\}$. So theoretically, we can compute the integral of, say, $f : \mathbb{N} \to [0, \infty]$ defined by $f(n) = n$ for all $n \in \mathbb{N}$ by finding all the nonnegative simple functions (with domain $\mathbb{N}$) that are less than or equal to f for all inputs $n \in \mathbb{N}$, then applying I to each function and finding the least upper bound. As an exercise, try computing an integral or two using this method.

We refer the reader to definition 4.7 in [2] for the definition of the integral of a measurable function $f : \Omega \to [-\infty, \infty]$. We do, however, present a bit of notation from definition 4.7 [2], as we will need this language in the definition of expected value.

Notation 4.2 ($\mathcal{L}^1(\Omega, \mathcal{A}, \mu)$)**.** *Let* $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$. *We say*

$$\mathcal{L}^1(\Omega, \mathcal{A}, \mu) := \left\{ f : \Omega \to \overline{\mathbb{R}} : f \textit{ is measurable and } \int |f| d\mu < \infty \right\}.$$

As will be seen, the definition of expected value is strikingly similar to the definition given in MAT340:

Definition 4.4 (Expected Value)**.** *Let* $(\Omega, \mathcal{A}, \mathbb{P})$ *be a probability space. If* $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, *then* X *is called integrable and we call*

$$\mathbf{E}[X] := \int X d\mathbb{P}$$

the expected value or mean of X. *(Definition 5.1 in* [2]*)*

As an exercise, visit the full definition of the integral in [2] and write the expected value of a random variable X as the supremum of a set (in the style of definition 4.3).

The definition of conditional expectation below appears in disconnect from the definition given in MAT340, but do not despair, for the following definition will provide comfort.

Definition 4.5 (conditional expectation)**.** *Let* $\mathcal{F} \subset \mathcal{A}$ *be a sub-*σ*-algebra and* $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, *where* $(\Omega, \mathcal{A}, \mathbb{P})$ *is a probability space. A random variable* Y *is called a conditional expectation of* X *given* $\mathcal{F}$, *symbolically* $\mathbf{E}[X \mid \mathcal{F}] := Y$, *if:*

(i) Y *is* $\mathcal{F}$*-measurable.*
(ii) For any $A \in \mathcal{F}$, *we have* $\mathbf{E}[X\mathbb{1}_A] = \mathbf{E}[Y\mathbb{1}_A]$.

For $B \in \mathcal{A}$, $\mathbb{P}[B \mid \mathcal{F}] := \mathbf{E}[\mathbb{1}_B \mid \mathcal{F}]$ *is called a conditional probability of* B *given the* σ*-algebra* $\mathcal{F}$. *(Definition 8.11 in* [2]*)*

Definition 4.6. *If* Y *is a random variable and* $X \in \mathcal{L}^1(\Omega, \mathcal{A}, \mathbb{P})$, *then we define* $\mathbf{E}[X \mid Y] := \mathbf{E}[X \mid \sigma(Y)]$. *(Definition 8.13 in* [2]*)*

Intuitively, we can think of a σ-algebra as providing information that might influence the predicted value of a random variable. For instance, $\mathbf{E}[X \mid \sigma(X)] = X$ intuitively because $\sigma(X)$ has all the information about X, so its true value has already been revealed.

5. Martingales

This section is devoted to defining a martingale, which is the essential aspect in the hypothesis of Azuma's inequality. Stochastic processes are those objects that we characterize as martingales, and to define stochastic processes, we must define the space on which they exist. Readers who are not familiar with real anaylsis should feel free to skip to definition 5.5. However, the material leading up to definition 5.5 serves as a great taste of the mathematical theory that underlies the objects seen in MAT340.

An advanced concept from real analysis is needed from the outset. A metric can be understood as a generalized notion of a distance function, and a metric space considers the set of objects to which distance is applied.

Definition 5.1 (metric space, metric). *A metric space (X, d) is a set X of objects (called points), together with a distance function or metric $d : X \times X \to [0, \infty)$, which associates to each pair x, y of points in X a non-negative real number $d(x, y) \geq 0$. Furthermore, the metric must satisfy the following four axioms:*

(a) For any $x \in X$, we have $d(x, x) = 0$.
(b) For any distinct $x, y \in X$, we have $d(x, y) > 0$.
(c) For any $x, y \in X$, we have $d(x, y) = d(x, y)$.
(d) For any $x, y, z \in X$, we have $d(x, z) \leq d(x, y) + d(y, z)$.

(definition 1.1.2 in [5]*)*

Recall from real analysis that a Cauchy sequence is one whose terms remain arbitrarily close as the index of the sequence approaches infinity (informal definition, of course).

Definition 5.2 (complete metric). *A metric d on X is said to be complete if any Cauchy sequence with respect to d converges in X.* [2]

For an example of a metric that is not complete, consider the metric $d(x, y) = |x - y|$ where $x, y \in \mathbb{Q}$, the set of rational numbers. Then the Cauchy sequence $3, 3.1, 3.14, 3.141, 3.1415, \cdots$ converges to $\pi \notin \mathbb{Q}$ [5]. Hence, $(d, \mathbb{Q})$ is not complete. However, $(d, \mathbb{R})$ is complete. Completeness can be thought of intuitively as lacking any holes. Note that complete spaces are useful in probability because many random variables take on real values such as π.

Definition 5.3 (separable). *A metric space is said to be separable if it has a countably dense subset.* [2]

For example, by a dense subset A in $\mathbb{R}$, we mean that if $x, y \in \mathbb{R}$ with $x < y$, then there exists $z \in A$ such that $x < z < y$. For instance, $\mathbb{Q}$ is a countably dense subset in $\mathbb{R}$.

Definition 5.4 (Polish Space). *A topological space (E, τ) is called a Polish Space if it is separable and if there exists a complete metric that induces the topology τ. (Definition 13.1 in* [2]*)*

Readers with background in elementary real analysis will recall how with the *usual* metric $d(x, y) = |x - y|$, one can define open sets. By a topology induced by a metric, we mean that we let the open sets of the topological space be precisely those sets that are open under the definition of the metric. A quote from [2] is motivating: "in practice, all spaces that are of importance in probability theory are Polish spaces."

For readers with interest in working as quantitative analysts on Wall Street, the following definition ought to become very familiar:

Definition 5.5 (stochastic process). *Let (E, τ) be a Polish space with Borel σ-algebra $\mathcal{E}$. Let $I \in \mathbb{R}$. A set of random variables $X = (X_t, t \in I)$(on $(\Omega, \mathcal{F}, \mathbb{P})$) with values in $(E, \mathcal{E})$ is called a stochastic process with index set (or time set) I and range E. (Definition 9.1 in* [2]*)*

Example 1. Let $I = \mathbb{N}$, where $i \in I$ denotes the number of days that have passed since January 1, 2020. Let X_i be the price of the most expensive stock on the NYSE at the end of day i. Then conceivably $(X_t, t \in I)$ is a stochastic process with values in $(\mathbb{R}, \mathcal{B}(R))$.

The following definition is crucial for understanding the definition of a martingale:

Definition 5.6 (generated σ-algebra). *Let $(\Omega', \mathcal{A}')$ be a measurable space and let Ω be a nonempty set. Let $X : \Omega \to \Omega'$ be a map. The inverse image*

$$X^{-1}(\mathcal{A}') := \{X^{-1}(A') : A' \in \mathcal{A}'\}$$

is the smallest σ-algebra with respect to which X is measurable. We say that $\sigma(X) := X^{-1}(\mathcal{A}')$ is the σ-algebra on Ω that is generated by X.

In considering σ-algebras that are generated by more than one map, we have the following: Let Ω be a nonempty set. Let I be an arbitrary index set. For any $i \in I$, let $(\Omega_i, \mathcal{A}_i)$ be a measurable space and let $X_i : \Omega \to \Omega_i$ be an arbitrary map. Then

$$\sigma(X_i, i \in I) := \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right) = \sigma\left(\bigcup_{i \in I} X_i^{-1}(\mathcal{A}_i)\right)$$

is called the σ-algebra on Ω that is generated by $(X_i, i \in I)$. This is the smallest σ-algebra with respect to which all X_i are measurable. (Theorem 1.78, Definition 1.79 in [2]*)*

Definition 5.7 (filtration). *Assume $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Let $\mathbb{F} = (\mathcal{F}_t, t \in I)$ be a family of σ-algebras with $\mathcal{F}_t \subset \mathcal{F}$ for all $t \in I$. We call $\mathbb{F}$ a filtration if $\mathcal{F}_s \subset \mathcal{F}_t$ for all $s, t \in I$ with $s \leq t$. (Definition 9.9 in* [2]*)*

Notice that a filtration on a countable index set I makes a sequence of σ-algebras $\mathcal{F}_i$ such that $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \cdots$.

Definition 5.8 (adapted). *A stochastic process $X = (X_t, t \in I)$ is called adapted to the filtration $\mathbb{F}$ if X_t is $\mathcal{F}_t$-measurable for all $t \in I$. If $\mathcal{F}_t = \sigma(X_s, s \leq t)$ for all $t \in I$, then we denote by $\mathbb{F} = \sigma(X)$ the filtration that is generated by X. (Definition 9.10 in* [2]*)*

For example, we can consider a stochastic process $X_1, X_2, \cdots$ together with a filtration $\mathcal{F}_1, \mathcal{F}_2, \cdots$ such that X_1 is $\mathcal{F}_1$-measurable (recall definition 3.4) and so forth. We now arrive at a famous term in probability theory:

Definition 5.9 (martingale). *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, let $I \subset \mathbb{R}$, and let $\mathbb{F}$ be a filtration. Let $X = (X_t)_{t \in I}$ be a real-valued, adapted stochastic process with $E[|X_t|] < \infty$ for all $t \in I$. Then X is called (with respect to $\mathbb{F}$) a* ***martingale*** *if $E[X_t \mid \mathcal{F}_s] = X_s$ for all $s, t \in I$ with $t > s$. (Definition 9.24 in* [2]*)*

Curious readers will be amused by a google search of the term, as it originates in gambling.

6. Azuma's Inequality

At long last:

Theorem 1 (Azuma's Inequality). *If $X = (M_n)_{n \in \mathbb{N}_0}$ is a martingale with respect to the filtration $\mathbb{F} = \sigma(X)$ with $M_0 = 0$ and if there is a sequence $(c_k)_{k \in \mathbb{N}}$ of nonnegative numbers with $|M_n - M_{n-1}| \leq c_n$ for all $n \in \mathbb{N}$, then for all $\lambda \geq 0$ we have*

$$\mathbb{P}[|M_n| \geq \lambda] \leq 2 \exp\left(-\frac{\lambda^2}{2 \sum_{k=1}^{n} {c_k}^2}\right).$$

The rest of the article is devoted to the proof of Azuma's inequality. We base this proof off of the one found in [3].

Definition 6.1 (convex function). *Let f be defined on an interval I. If for all $x_1, x_2 \in I$ and $\alpha \in [0,1]$ the inequality*

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$$

is satisfied, we say f is convex on I. (Definition 7.33 in [1]*)*

Lemma 2. *We will show that for any $x \in [-1,1]$ and $t > 0$, then*

$$e^{tx} \leq \frac{1}{2}(1+x)e^{t} + \frac{1}{2}(1-x)e^{-t}.$$

Proof. Let $x \in [-1,1]$, $t > 0$, $x_1 = -1$, and $x_2 = 1$. Further, let $f(y) = e^{ty}$ for all $y \in [-1,1]$. From calculus, we know that f is a convex function on $[-1,1]$. Hence, by definition 6.1,

$$\exp(t \cdot [\alpha(-1) + (1-\alpha)(1)]) \leq \alpha e^{-t} + (1-\alpha)e^{t}$$

and

$$\exp(t \cdot (1-2\alpha)) \leq \alpha e^{-t} + (1-\alpha)e^{t} \tag{1}$$

for all $\alpha \in [0,1]$. Now let $\alpha = \frac{1}{2}(1-x)$. Let us verify that $0 \leq \alpha \leq 1$. We have

$$\begin{aligned}
&-1 \leq x \leq 1 && \\
&\frac{-1}{2} \leq \frac{x}{2} \leq \frac{1}{2} && (\text{add } \frac{1}{2}) \\
&\frac{1}{2} \geq -\frac{x}{2} \geq -\frac{1}{2} && (\text{mulitply by } -1) \\
&1 \geq \frac{1}{2} - \frac{x}{2} \geq 0 && (\text{add } \frac{1}{2}) \\
&1 \geq \alpha \geq 0, &&
\end{aligned}$$

as desired. Substituting for α in equation (1) tells us that

$$\begin{aligned}
\exp\left(t \cdot \left(1 - 2\left(\frac{1}{2}(1-x)\right)\right)\right) &\leq \frac{1}{2}(1-x)e^{-t} + (1 - \frac{1}{2}(1-x))e^{t} \\
e^{tx} &\leq \frac{1}{2}(1-x)e^{-t} + \left(\frac{1}{2} + \frac{1}{2}x\right)e^{t} \\
e^{tx} &\leq \frac{1}{2}(1-x)e^{-t} + \frac{1}{2}(1+x)e^{t},
\end{aligned}$$

as desired. □

Lemma 3. *Let Y be a random variable such that $Y \in [-1,1]$ and $\mathbf{E}[Y \mid \mathcal{F}] = 0$. Then for any $t \geq 0$, we have that $\mathbf{E}[e^{tY} \mid \mathcal{F}] \leq e^{\frac{t^2}{2}}$.*

Proof. Fix $t \geq 0$. By Lemma 2 we have that $e^{tY} \leq \frac{1}{2}(1+Y)e^t + \frac{1}{2}(1-Y)e^{-t}$. Then by monotonicity of conditional expectation we have

$$\begin{aligned}
\mathbf{E}[e^{tY} \mid \mathcal{F}] &\leq \mathbf{E}[\frac{1}{2}(1+Y)e^t + \frac{1}{2}(1-Y)e^{-t} \mid \mathcal{F}] \\
&= \mathbf{E}[\frac{1}{2}(1+Y)e^t \mid \mathcal{F}] + \mathbf{E}[\frac{1}{2}(1-Y)e^{-t} \mid \mathcal{F}] \\
&= \mathbf{E}[\frac{1}{2}e^t \mid \mathcal{F}] + \mathbf{E}[\frac{1}{2}Ye^t \mid \mathcal{F}] + \mathbf{E}[\frac{1}{2}e^{-t} \mid \mathcal{F}] - \mathbf{E}[\frac{1}{2}Ye^{-t} \mid \mathcal{F}] \\
&= \frac{1}{2}e^t + \frac{1}{2}e^t\mathbf{E}[Y \mid \mathcal{F}] + \frac{1}{2}e^{-t} - \frac{1}{2}e^{-t}\mathbf{E}[Y \mid \mathcal{F}] \\
&= \frac{1}{2}e^t + \frac{1}{2}e^{-t} \\
&= \cosh(t) \\
&= \sum_{i=0}^{\infty} \frac{t^{2i}}{(2i)!} \leq \sum_{i=0}^{\infty} \frac{t^{2i}}{2^i(i)!} = \sum_{i=0}^{\infty} \frac{\left(\frac{t^2}{2}\right)^i}{i!} = e^{t^2/2}.
\end{aligned}$$

□

Fact 6.1 (Markov's Inequality (taken from [2]))**.** *Let X be a real random variable and let $f : [0,\infty) \to [0,\infty)$ be monotone increasing. Then for any $\mathcal{E} > 0$ with $f(\mathcal{E}) > 0$, the Markov inequality holds,*

$$\mathbb{P}[|X| \geq \mathcal{E}] \leq \frac{\mathbf{E}[f(|x|)]}{f(\mathcal{E})}.$$

Proof of Azuma's Inequality:

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space. Let $X = (M_n)_{n\in\mathbb{N}_0}$ be a martingale with respect to the filtration $\mathbb{F} = \sigma(X)$ with $M_0 = 0$. Further, let $(c_k)_{k\in\mathbb{N}}$ be a sequence of nonnegative numbers with $|M_n - M_{n-1}| \leq c_n$ for all $n \in \mathbb{N}$, and let $\lambda \geq 0$. Note that all M_n have values in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

Before getting into the heart of the proof, it will be helpful to note some aspects of the martingale X and its filtration. We show that since $M_0(\omega) = 0$ for all $\omega \in \Omega$, then M_0 is $\{\emptyset, \Omega\}$-measurable. Note that $\{\emptyset, \Omega\}$ is a σ-algebra. Let $B \in \mathcal{B}(\mathbb{R})$. If $0 \in B$, then $M_0^{-1}(B) = \Omega \in \{\emptyset, \Omega\}$. If $0 \notin B$, then $M_0^{-1}(B) = \emptyset \in \{\emptyset, \Omega\}$, since no value in the domain Ω is mapped to a nonzero real number. By definition 3.4, M_0 is $\{\emptyset, \Omega\}$-measurable. Since $\mathcal{F}_0 = \sigma(M_s, s \leq 0)$, then $\mathcal{F}_0 = \sigma(M_0) = M_0^{-1}(\mathcal{B}(\mathbb{R})) = \{\emptyset, \Omega\}$, by definitions 5.8 and 5.6 for the first and second equalities. Further, $\{\emptyset, \Omega\} = \mathcal{F}_0 \subset \mathcal{F}_1 = \sigma(\{M_0, M_1\}) \subset \cdots \subset \mathcal{F}_n = \sigma(\{M_0, \cdots, M_n\}) \subset \cdots$, characterizes $\mathbb{F}$. Consequently, M_n is $\mathcal{F}_n$-measurable for all n. Thus, we will denote $M_n : (\Omega, \mathcal{F}_n) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for all n.

Define $Y_n : (\Omega, \mathcal{F}_n) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by $Y_n(\omega) = (M_n - M_{n-1})(\omega)$ for all $\omega \in \Omega$. That is, $Y_1 = M_1 - M_0$, $Y_2 = M_2 - M_1$, and so forth. Thus, $|Y_n| \leq c_n$ by the given hypothesis. We will show that $\mathbf{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$. Observe:

$$\begin{aligned}
\mathbf{E}[Y_n \mid \mathcal{F}_{n-1}] &= \mathbf{E}[M_n - M_{n-1} \mid \mathcal{F}_{n-1}] \\
&= \mathbf{E}[M_n \mid \mathcal{F}_{n-1}] - \mathbf{E}[M_{n-1} \mid \mathcal{F}_{n-1}] \\
&= M_{n-1} - \mathbf{E}[M_{n-1} \mid \mathcal{F}_{n-1}] && \text{(definition of martingale)} \\
&= M_{n-1} - \mathbf{E}[M_{n-1} \mid \sigma(\{M_{n-1}, \cdots, M_0\})] && \text{(since } \mathbb{F} = \sigma(X)) \\
&= M_{n-1} - M_{n-1} && \text{(see Thm 8.14 in [2])} \\
&= 0.
\end{aligned}$$

Next, let $t > 0$. Let $f : [0, \infty) \to [0, \infty)$ be defined by $f(x) = e^{tx}$ for all $x \in [0, \infty)$. Thus, f is monotone increasing. Then by Markov's inequality, given $\lambda > 0$, it follows that

$$\mathbb{P}[|M_n - M_0| \geq \lambda] \leq \frac{\mathbf{E}[e^{t|M_n - M_0|}]}{e^{t\lambda}}.$$

Then write $M_n = Y_n + M_{n-1}$ and substitute:

$$\begin{aligned}
\mathbb{P}[|M_n - M_0| \geq \lambda] &\leq \frac{\mathbf{E}[e^{t|Y_n + M_{n-1} - M_0|}]}{e^{t\lambda}} \\
&\leq \frac{\mathbf{E}[e^{t(|Y_n| + |M_{n-1} - M_0|)}]}{e^{t\lambda}} && \text{(Triangle inequality)} \\
&= e^{-t\lambda}\mathbf{E}[\mathbf{E}[e^{t|Y_n|}e^{t|M_{n-1} - M_0|} \mid \mathcal{F}_{n-1}]] && \text{(Tower property, see Thm 8.14 [2]).}
\end{aligned}$$

Next, notice that $e^{t|M_{n-1} - M_0|}$ is $\mathcal{F}_{n-1}$-measurable, so it can be factored out of the inner expectation, by Theorem 8.14 [2]:

$$\mathbf{E}[e^{t|Y_n|}e^{t|M_{n-1} - M_0|} \mid \mathcal{F}_{n-1}] = e^{t|M_{n-1} - M_0|}\mathbf{E}[e^{t|Y_n|} \mid \mathcal{F}_{n-1}].$$

Furthermore, since $\mathbf{E}[Y_n \mid \mathcal{F}_{n-1}] = 0$, we infer that $\mathbf{E}[\frac{|Y_n|}{c_n} \mid \mathcal{F}_{n-1}] = 0$. In addition to the fact that $\frac{|Y_n|}{c_n} \in [-1, 1]$, we can apply Lemma 3 to obtain

$$\begin{aligned}
\mathbf{E}[e^{t|Y_n|} \mid \mathcal{F}_{n-1}] &= \mathbf{E}[e^{tc_n \frac{|Y_n|}{c_n}} \mid \mathcal{F}_{n-1}] \\
&\leq e^{t^2 c_n^2/2}.
\end{aligned}$$

Thus, by substitution,

$$\begin{aligned}
\mathbb{P}[|M_n - M_0| \geq \lambda] &\leq e^{-t\lambda}\mathbf{E}[e^{t|M_{n-1} - M_0|}\mathbf{E}[e^{t|Y_n|} \mid \mathcal{F}_{n-1}]] \\
&\leq e^{-t\lambda}e^{t^2 c_n^2/2}\mathbf{E}[e^{t|M_{n-1} - M_0|}].
\end{aligned}$$

Notice that the term $\mathbf{E}[e^{t|M_{n-1} - M_0|}]$ looks like the term $\mathbf{E}[e^{t|M_n - M_0|}]$ that we started with. Hence, proceed recursively until the subscript on M has reached 0. Thus, we obtain

$$\mathbb{P}[|M_n - M_0| \geq \lambda] \leq \exp(-t\lambda + t^2 \sum_{i=1}^{n} \frac{c_i^2}{2}),$$

for all $t > 0$. Now, to make this bound as effective as possible, we want to optimize t so that $\exp(-t\lambda + t^2 \sum_{i=1}^{n} \frac{c_i^2}{2})$ is minimal. It turns out that choosing $t = \frac{\lambda}{\sum_{i=1}^{n} c_i^2}$ achieves this. It will be left as an exercise to show why. Thus, with $M_0 = 0$, we have obtained Azuma's inequality:

$$\mathbb{P}[|M_n| \geq \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2\sum_{k=1}^{n} {c_k}^2}\right).$$

□

References

[1] A. M. Bruckner, B. S. Thomson, J. B. Bruckner, *Elementary Real Analysis*, Second Edition, 2008.

[2] Achim Klenke, *Probability Theory*, Springer, Second Edition, 2014.

[3] Alistair Sinclair, “Lecture 18: March 19: Martingales”, CSC271 Randomness and Computation, Spring 2020.

[4] Terrence Tao, *Analysis 1*, Hindustan Book Agency (India), 2017.

[5] Terrence Tao, *Analysis 2*, Hindustan Book Agency (India), Third Edition, 2017.