CHAPTER 2

Brownian motion, and an Introduction to Modern Probability

1. Scaling limit of random walks.

Our first goal is to understand *Brownian motion*, which is used to model "noisy fluctuations" of stocks, and various other objects. This is named after the botanist Robert Brown, who observed that the microscopic movement of pollen grains appears random. Intuitively, Brownian motion can be thought of as a process that performs a random walk in continuous time.

We begin by describing Brownian motion as the scaling limit of discrete random walks. Let ξ_1, ξ_2, ..., be a sequence of i.i.d. random variables which take on the values ± 1 with probability $1/2$. Define the time interpolated random walk $S(t)$ by setting $S(0) = 0$, and

$$S(t) = S(n) + (t - n)\xi_{n+1} \quad \text{when } t \in (n, n+1]\,. \tag{1.1}$$

Note $S(n) = \sum_1^n \xi_i$, and so at integer times S is simply a symmetric random walk with step size 1.

Our aim now is to rescale S so that it takes a random step at shorter and shorter time intervals, and then take the limit. In order to get a meaningful limit, we will have to compensate by also scaling the step size. Let $\varepsilon > 0$ and define

$$S_\varepsilon(t) = \alpha_\varepsilon S\Big(\frac{t}{\varepsilon}\Big)\,, \tag{1.2}$$

where α_ε will be chosen below in a manner that ensures convergence of $S_\varepsilon(t)$ as $\varepsilon \to 0$. Note that S_ε now takes a random step of size α_ε after every ε time units.

To choose α_ε, we compute the variance of S_ε. Note first

$$\operatorname{Var} S(t) = \lfloor t \rfloor + (t - \lfloor t \rfloor)^2,$$

and[1] consequently

$$\operatorname{Var} S_\varepsilon(t) = \alpha_\varepsilon^2 \Big(\Big\lfloor \frac{t}{\varepsilon} \Big\rfloor + \Big(\frac{t}{\varepsilon} - \Big\lfloor \frac{t}{\varepsilon} \Big\rfloor \Big)^2 \Big)\,.$$

In order to get a "nice limit" of S_ε as $\varepsilon \to 0$, one would at least expect that $\operatorname{Var} S_\varepsilon(t)$ converges as $\varepsilon \to 0$. From the above, we see that choosing

$$\alpha_\varepsilon = \sqrt{\varepsilon}$$

immediately implies

$$\lim_{\varepsilon \to 0} \operatorname{Var} S_\varepsilon(t) = t\,.$$

Theorem 1.1. *The processes $S_\varepsilon(t) \overset{\text{def}}{=} \sqrt{\varepsilon} S(t/\varepsilon)$ "converge" as $\varepsilon \to 0$. The limiting process, usually denoted by W, is called a (standard, one dimensional)* Brownian motion.

The proof of this theorem uses many tools from the modern theory of probability, and is beyond the scope of this course. The important thing to take away from this is that Brownian motion can be well approximated by a random walk that takes steps of variance ε on a time interval of size ε.

While the above construction provides good intuition as to what Brownian motion actually is, the scaling limit it is a somewhat unwieldy object to work with. We instead introduce an intrinsic characterization of Brownian motion, and we will shortly see that is both useful and mathematically convenient.

Definition 1.2. A Brownian motion is a continuous process that has stationary independent increments.

Let us briefly explain the terms appearing in the above definition.

(1) A *process* (aka stochastic process) is simply a collection of random variables $\{X(t) \mid 0 \leqslant t < \infty\}$. The index t usually represents time, and the process is often written as $\{X_t \mid 0 \leqslant t < \infty\}$ instead.
(2) A *trajectory* (aka sample path) of a process is the outcome of one particular realization each of the random variables $X(t)$ viewed as function of time.
(3) A process is called *continuous* if each of the trajectories are continuous. That is, for every $t > 0$ we have

$$\lim_{s \to t} X(s) = X(t)\,. \tag{1.3}$$

(4) An process is said to have *stationary increments* if for every $h \geqslant 0$, the distribution of $X(t+h) - X(t)$ does not depend on t.
(5) A process is said to have *independent increments* if for every finite sequence of times $0 \leqslant t_0 < t_1 \cdots < t_N$, the random variables $X(t_0)$, $X(t_1) - X(t_0)$, $X(t_2) - X(t_1)$, ..., $X(t_N) - X(t_{N-1})$ are all jointly independent.

For the process S in (1.1), note that for $n \in \mathbb{N}$, $S(n+1) - S(n) = X_{n+1}$ whose distribution *does not* depend on n as the variables $\{\xi_i\}$ were chosen to be independent and identically distributed. Similarly, $S(n+k) - S(n) = \sum_{n+1}^{n+k} \xi_i$ which has the same distribution as $\sum_1^k \xi_i$ and is independent of n.

However, if $t \in \mathbb{R}$ and is not necessarily an integer, $S(t+k) - S(t)$ will in general depend on t. So the process S (and also S_ε) do not have stationary (or independent) increments.

We claim, that the limiting process W does have stationary, independent, *normally distributed* increments. Suppose for some fixed $\varepsilon > 0$, both s and t are multiples of ε. In this case

$$S_\varepsilon(t) - S_\varepsilon(s) \sim \sqrt{\varepsilon} \sum_{i=1}^{\lfloor t-s \rfloor/\varepsilon} \xi_i \xrightarrow{\varepsilon \to 0} N(0, t-s)\,,$$

by the central limit theorem. If s, t aren't multiples of ε as we will have in general, the first equality above is true up to a remainder which can easily be shown to vanish.

[1] Here $\lfloor x \rfloor$ denotes the greatest integer smaller than x. That is, $\lfloor x \rfloor = \max\{n \in \mathbb{Z} \mid n \leqslant x\}$.

The above heuristic argument suggests that the limiting process W (from Theorem 1.1) satisfies $W(t) - W(s) \sim N(0, t-s)$. This certainly has independent increments since $W(t+h) - W(t) \sim N(0,h)$ which is independent of t. Moreover, this also suggests that Brownian motion can be *equivalently* characterized as follows.

DEFINITION 1.3. A Brownian motion is a *continuous process* W such that:

1. W has independent increments, and
2. For $s < t$, $W(t) - W(s) \sim N(0, \sigma^2(t-s))$.

REMARK 1.4. A *standard* (one dimensional) Brownian motion is one for which $W(0) = 0$ and $\sigma = 1$.

2. A brief review of probability

In modern probability we usually start with a *probability triple* $(\Omega, \mathcal{G}, \boldsymbol{P})$.

1. Ω is a non-empty set called the *sample space.*
2. $\mathcal{G}$ is a σ-algebra. This is a non-empty collection of events (subsets of Ω) of which the probability is known.
3. $\boldsymbol{P}$ is a *probability measure.* For any event $A \in \mathcal{G}$, $\boldsymbol{P}(A)$ represents the probability of the event A occurring.

A subtle, but important, point in this framework is the in most case $\mathcal{G}$ is usually not the collection of all subsets of Ω, but only a collection of some subsets of Ω. In fact, in most interesting examples, it is *impossible* to define the probability of arbitrary subsets of Ω consistently (i.e. in a manner that satisfies the required properties listed below), and we thus restrict ourselves to only talking about probabilities of elements of elements of the σ-algebra $\mathcal{G}$.

In order to be a probability space, the triple $(\Omega, \mathcal{G}, \boldsymbol{P})$ is required to satisfy certain properties. First the σ-algebra $\mathcal{G}$ must satisfy the following:

1. It must be closed under compliments. That is, if $A \in \mathcal{G}$, then $A^c \in \mathcal{G}$.
2. It must be closed under *countable* unions. That us, if A_1, A_2, ... are all elements of $\mathcal{G}$, then the union $\cup_1^\infty A_i$ is also an element of $\mathcal{G}$.

The precise mathematical definition of a σ-algebra is simply a non-empty collection of sets that satisfies the above two properties. Of course, if $\mathcal{G}$ satisfies the above properties then one can quickly deduce the following:

3. The empty set $\emptyset$ and the whole space Ω are elements of $\mathcal{G}$.
4. If A_1, A_2, ... are all elements of $\mathcal{G}$, then the intersection $\cap_1^\infty A_i$ is also an element of $\mathcal{G}$.
5. If A, B are events in $\mathcal{G}$, then $A - B$ is also an event in $\mathcal{G}$.

The reason for requiring the above properties is that we expect $\mathcal{G}$ is the collection of events of which the probability is known (or of which the probability can be deduced by performing repeated trials of some experiment). If you can deduce the probability of an event A, you should certainly be able to deduce the probability of A^c. Similarly, if you can deduce the probabilities of A, B, you should be able to deduce the probability of $A \cup B$ and $A \cap B$. The possibly surprising point above is that we don't require that $\mathcal{G}$ be closed under finite unions, but we require it is closed under *countable* unions. The reason for this is that we would like our framework to allow us to perform repeated trials of an experiment and take limits.

Next, we turn our attention to the probability measure $\boldsymbol{P}$. We require that $\boldsymbol{P}$ satisfies the following properties:

1. For each $A \in \mathcal{G}$, $\boldsymbol{P}(A) \in [0,1]$. Moreover, $\boldsymbol{P}(\emptyset) = 0$, and $\boldsymbol{P}(\Omega) = 1$.
2. *(Countable additivity)* Given *pairwise disjoint* events $A_1, A_2, \dots \in \mathcal{G}$, we have
$$\boldsymbol{P}\Big(\bigcup_{i=1}^\infty A_i\Big) = \sum_{i=1}^\infty \boldsymbol{P}(A_i)\,.$$

The above two properties are precisely the formal definition of a probability measure. Recall that in probability we require that the probability of mutually exclusive events add. The second property above is a generalization of this to countably many events.

Using the above properties, one can quickly verify that $\boldsymbol{P}$ also satisfies the following properties:

1. $\boldsymbol{P}(A^c) = 1 - \boldsymbol{P}(A)$. More generally, if $A, B \in \mathcal{G}$ with $A \subseteq B$, then $\boldsymbol{P}(B - A) = \boldsymbol{P}(B) - \boldsymbol{P}(A)$.
2. If $A_1 \subseteq A_2 \subseteq A_3 \cdots$ and each $A_i \in \mathcal{G}$ then $\boldsymbol{P}(\cup A_i) = \lim_{n\to\infty} \boldsymbol{P}(A_n)$.
3. If $A_1 \supseteq A_2 \supseteq A_3 \cdots$ and each $A_i \in \mathcal{G}$ then $\boldsymbol{P}(\cap A_i) = \lim_{n\to\infty} \boldsymbol{P}(A_n)$.

We now describe *random variables* in the above context. In discrete probability, random variables are usually just real valued functions defined on the sample space. In our context, however, we need to be a bit more careful. If X is a random variable, then one should always be able to assign probabilities to questions such as *"Is X positive?"* or *"Does X belong to the interval $(0,1)$?"*.

If X is simply a function from Ω to $\mathbb{R}$, then to compute the probability that X is positive, we should define $A = \{\omega \in \Omega \mid X(\omega) > 0\}$, and then compute $\boldsymbol{P}(A)$. This, however, is only possible if $A \in \mathcal{G}$; and since $\mathcal{G}$ *is usually not* the entire power set of Ω, we should take care to ensure that all questions we might ask about the random variable X can be answered by only computing probabilities of events in $\mathcal{G}$, and not arbitrary subsets of Ω. For this reason, we define random variables as follows.

DEFINITION 2.1. A *random variable* is a *$\mathcal{G}$-measurable function* $X : \Omega \to \mathbb{R}$. That is, a random variable is a function $X\colon \Omega \to \mathbb{R}$ such that for every $\alpha \in \mathbb{R}$, the set $\{\omega \in \Omega \mid X(\omega) \leqslant \alpha\}$ is guaranteed to be an element of $\mathcal{G}$. (Such functions are also called *$\mathcal{G}$-measurable*, *measurable with respect to* $\mathcal{G}$, or simply *measurable* if the σ-algebra in question is clear from the context.)

REMARK 2.2. The argument ω is *always* suppressed when writing random variables. That is, the event $\{\omega \in \Omega \mid X(\omega) \leqslant \alpha\}$ is simply written as $\{X \leqslant \alpha\}$.

REMARK 2.3. Note for any random variable, $\{X > \alpha\} = \{X \leqslant \alpha\}^c$ which must also belong to $\mathcal{G}$ since $\mathcal{G}$ is closed under complements. One can check that for every $\alpha < \beta \in \mathbb{R}$ the events $\{X < \alpha\}$, $\{X \geqslant \alpha\}$, $\{X > \alpha\}$, $\{X \in (\alpha, \beta)\}$, $\{X \in [\alpha, \beta)\}$, $\{X \in (\alpha, \beta]\}$ and $\{X \in (\alpha, \beta)\}$ are all also elements of $\mathcal{G}$.

Thus to (for instance) compute the chance that X lies strictly between two real numbers α and β, we consider the event $\{X \in (\alpha, \beta)\}$. By Remark 2.3 this is guaranteed to be an element of $\mathcal{G}$, and thus we can compute the probability of it using $\boldsymbol{P}$. Hence, the quantity $\boldsymbol{P}(\{X \in (\alpha, \beta)\})$ is mathematically well defined, and represents the chance that the random variable X takes values in the interval (α, β). For brevity, we almost always omit the outermost curly braces and write $\boldsymbol{P}(X \in (\alpha, \beta))$ for $\boldsymbol{P}(\{X \in (\alpha, \beta)\})$.

Remark 2.4. One can check that if X, Y are random variables then so are $X \pm Y$, XY, X/Y (when defined), $|X|$, $X \wedge Y$ and $X \vee Y$. In fact if $f : \mathbb{R} \to \mathbb{R}$ is any reasonably nice (more precisely, a Borel measurable) function, $f(X)$ is also a random variable.

Example 2.5. Given $A \subseteq \Omega$, define *indicator function of A* by

$$\mathbf{1}_A(\omega) \stackrel{\text{def}}{=} \begin{cases} 1 & \omega \in A\,, \\ 0 & \omega \notin A\,. \end{cases}$$

One can check that $\mathbf{1}_A$ is a ($\mathcal{G}$-measurable) random variable if and only if $A \in \mathcal{G}$.

Example 2.6. For $M \in \mathbb{N}$, $i \in \{1, \dots, M\}$, $a_i \in \mathbb{R}$ and $A_i \in \mathcal{G}$ be such that $A_i \cap A_j = \emptyset$ for $i \neq j$, and define

$$X \stackrel{\text{def}}{=} \sum_{i=1}^{M} a_i \mathbf{1}_{A_i}\,. \tag{2.1}$$

Then X is a ($\mathcal{G}$-measurable) random variable. (Such variables are called *simple* random variables.)

The next important concept concerning random variables is that of *expectation*, which we assume the reader is familiar with in the discrete setting. In the measure theoretic framework, the expectation of a random variable is the *Lebesgue integral*, and is denoted by[2]

$$\boldsymbol{E}X \stackrel{\text{def}}{=} \int_\Omega X \, d\boldsymbol{P}\,.$$

The precise construction of the Lebesgue integral, however, is to lengthy to be presented here, and we only present a brief summary.

If a random variable X only takes on finitely many values $a_1, \dots a_n$, then the expectation of X is given by

$$\boldsymbol{E}X \stackrel{\text{def}}{=} \sum_{i=1}^{n} a_i \boldsymbol{P}(X = a_i)\,. \tag{2.2}$$

This means that for any *simple random variable* of the form (2.1), the expectation is given by (2.2). For general random variables (i.e. random variables that are not simple), we can compute by expressing them as a limit of simple random variables. Namely, we can compute $\boldsymbol{E}X$ by

$$\boldsymbol{E}X = \lim_{n \to \infty} \boldsymbol{E}\Big(\sum_{k=-n^2}^{n^2-1} \frac{k}{n} \mathbf{1}_{\{\frac{k}{n} \leqslant X < \frac{k+1}{n}\}} \Big) = \lim_{n \to \infty} \sum_{k=-n^2}^{n^2-1} \frac{k}{n} \boldsymbol{P}\Big(\frac{k}{n} \leqslant X < \frac{k+1}{n} \Big)\,,$$

for instance.

The above description, however, is only of theoretical importance and is not used to compute in practice. Here are a few computation rules and properties of expectations that will be useful later.

(1) *(Linearity)* If $\alpha \in \mathbb{R}$ and X, Y are random variables, then $\boldsymbol{E}(X + \alpha Y) = \boldsymbol{E}X + \alpha \boldsymbol{E}Y$.
(2) *(Positivity)* If $X \geqslant 0$ almost surely,[3] then $\boldsymbol{E}X \geqslant 0$. Moreover, if $X > 0$ almost surely, $\boldsymbol{E}X > 0$. Consequently, (using linearity) if $X \leqslant Y$ almost surely then $\boldsymbol{E}X \leqslant \boldsymbol{E}Y$.
(3) *(Layer Cake formula)* If $X \geqslant 0$ almost surely, then

$$\boldsymbol{E}X = \int_0^\infty \boldsymbol{P}(X \geqslant t)\, dt\,.$$

More generally, if φ is a increasing differentiable function with $\varphi(0) = 0$ then

$$\boldsymbol{E}\varphi(X) = \int_0^\infty \varphi'(t)\, \boldsymbol{P}(X \geqslant t)\, dt\,.$$

(4) *(Unconscious Statistician Formula)* If the probability density function of X is p, and f is any (Borel measurable) function, then

$$\boldsymbol{E}f(X) = \int_{-\infty}^{\infty} f(x) p(x)\, dx\,. \tag{2.3}$$

The proof of these properties goes beyond the scope of these notes. We do, however, make a few remarks. It turns out that the proof of positivity in this framework is immediate, however the proof of linearity is surprisingly not as straightforward as you would expect. While it is easy to verify linearity for simple random variables, for general random variables, the proof of linearity requires an approximation argument. The full proof of this involves either the *dominated* or *monotone* convergence theorem which guarantee $\lim \boldsymbol{E}X_n = \boldsymbol{E} \lim X_n$, under modest assumptions.

The layer cake formula can be proved by drawing a graph of X with Ω on the horizontal axis. Now $\boldsymbol{E}X$ should be the "area under the curve", which is usually computed by slicing the region into vertical strips and adding up the area of each strip. If, instead, you compute the area by slicing the region into *horizontal* strips, you get exactly the layer cake formula!

Finally, unconscious statistician formula might already be familiar to you. In fact, the reason for this somewhat unusual name is that many people use this result "unconsciously" treating it as the definition, without realizing it is in fact a theorem that requires proof. To elaborate further, introductory (non-measure theoretic) probability courses usually stipulate that if a random variable X has density p_X, then

$$\boldsymbol{E}X = \int_{-\infty}^{\infty} x p_X(x)\, dx\,.$$

[2] If $A \in \mathcal{G}$ we define

$$\int_A Y \, d\boldsymbol{P} \stackrel{\text{def}}{=} \boldsymbol{E}(\mathbf{1}_A Y)\,,$$

and when $A = \Omega$ we will often omit writing it.

[3] By $X \geqslant 0$ almost surely, we mean that $\boldsymbol{P}(X \geqslant 0) = 1$. More generally, we say an event occurs almost surely if the probability of it occurring is 1.

Thus if you set $Y = f(X)$ for some function f, we should have

$$\boldsymbol{E}Y = \int_{-\infty}^{\infty} y p_Y(y)\, dy\,.$$

If we could compute p_Y in terms of p_X and f, you could substitute it in the above formula, and obtain a formula for $\boldsymbol{E}Y$ in terms of p_X and f. Unfortunately, this isn't easy to do. Namely, if f isn't monotone, it isn't easy to write down p_Y in terms of p_X. It turns out, however, that even though we can't easily write down p_Y in terms of f and p_X, we can prove that $\boldsymbol{E}Y$ can be computed using (2.3).

Since discussing these results and proofs further at this stage will will lead us too far astray, we invite the curious to look them up in any standard measure theory book. The main point of this section was to introduce you to a framework which is capable of describing and studying the objects we will need for the remainder of the course.

We conclude this section by revisiting the notion of a *continuous process* defined in the previous section. Recall, our definition so far was that a process is simply a collection of random variables $\{X(t)\}_{t\geqslant 0}$, and a continuous process is a process whose trajectories are continuous. In our context, a process can now be thought of as a function

$$X\colon \Omega \times [0,\infty) \to \mathbb{R}\,.$$

For every fixed t, the function $\omega \mapsto X(\omega, t)$ is required to be a random variable (i.e. measurable with respect to $\mathcal{G}$). Since the ω is usually suppressed in probability, this random variable is simply denoted by $X(t)$.

The trajectory of X is now the slice of X for a fixed ω. Namely, for any fixed $\omega \in \Omega$, the function $t \mapsto X(\omega, t)$ is the trajectory of X. Saying a process has continuous trajectories means that for every $\omega \in \Omega$, the trajectory $t \mapsto X(\omega, t)$ is continuous as a function of t. Explicitly, this means for every $t \geqslant 0$ and $\omega \in \Omega$ we have

$$\lim_{s\to t} X(\omega, s) = X(\omega, s)\,.$$

Following our convention of "never writing ω", this is exactly (1.3) as we had before.

3. Independence of random variables

Recall two events A, B are independent if $P(A \mid B) = P(A)$. This is of course immediately implies the multiplication law:

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

The notion of independence for random variables requires that *every* event that is observable from one is necessarily independent of *every* event that is observable from the other.

For example, suppose X and Y are two random variables. For any $a, b \in \mathbb{R}$, the event $\{X \in (a,b)\}$ can be observed using the random variable X. Similarly, any $c \in \mathbb{R}$, the event $\{Y > c\}$ can be observed using the random variable Y. If X and Y were independent, then the events $\{X \in (a,b)\}$ would necessarily be independent of the event $\{Y > c\}$. Of course, this is just an example and you can write down all sorts of other events (e.g. $X^2 - e^X < 15$, or $\sin(Y+3) < .5$). No matter how you do it, if X and Y are independent, then any event observable from X alone must necessarily be independent of any event observable from Y alone.

Since the notion of "all events that can be observed from the random variable X" will be useful later, we denote it by $\sigma(X)$.

Definition 3.1. Let X be a random variable on $(\Omega, \mathcal{G}, \boldsymbol{P})$. We define *the σ-algebra generated by* X to be the σ algebra obtained by only using events that are observable using the random variable X.

One can mathematically prove that $\sigma(X)$ is generated by the events $\{X \leqslant \alpha\}$ for every $\alpha \in \mathbb{R}$. Namely, if a σ algebra contains the events $\{X \leqslant \alpha\}$ for every $\alpha \in \mathbb{R}$, then it must necessarily contain *all* events observable through the random variable X. In particular, it will contain events of the form $\{X \in [\alpha, \beta)\}$, $e^{X+1} < \sin X$, or any other complicated formula that you can write down.

As mentioned above, the σ-algebra $\sigma(X)$ represents all the information one can obtain by observing X. To illustrate this, consider the following example: A card is drawn from a shuffled deck, and you win a dollar if it is red, and lose one if it is black. Now the likely hood of drawing any particular card is $1/52$. However, if you are blindfolded and only told the outcome of the game, you have no way to determine that each gard is picked with probability $1/52$. The only thing you will be able to determine is that red cards are drawn as often as black ones.

This is captured by σ-algebra as follows. Let $\Omega = \{1, \dots, 52\}$ represent a deck of cards, $\mathcal{G} = \mathcal{P}(\Omega)$, and define $\boldsymbol{P}(A) = \operatorname{card}(A)/52$. Let $R = \{1, \dots 26\}$ represent the red cards, and $B = R^c$ represent the black cards. The outcome of the above game is now the random variable $X = \boldsymbol{1}_R - \boldsymbol{1}_B$, and you should check that $\sigma(X)$ is exactly $\{\emptyset, R, B, \Omega\}$.

With this, we can now revisit the notion of two random variables being independent.

Definition 3.2. We say the random variables $X_1, \dots, X_N$ are independent if for every $i \in \{1 \dots N\}$ and every $A_i \in \sigma(X_i)$ the events $A_1, \dots, A_N$ are independent.

Remark 3.3. Recall, A collection of events $A_1, \dots, A_N$ is said to be independent if *any* sub collection $\{A_{i_1}, \dots, A_{i_k}\}$ satisfies the multiplication law

$$\boldsymbol{P}\Bigl(\bigcap_{i=1}^{k} A_{i_k}\Bigr) = \prod_{i=1}^{d} \boldsymbol{P}(A_i)\,.$$

Note that this is a *stronger* condition than simply requiring

$$\boldsymbol{P}\bigl(A_1 \cap A_2 \cap \cdots \cap A_N\bigr) = \boldsymbol{P}(A_1)\,\boldsymbol{P}(A_2) \cdots \boldsymbol{P}(A_N)\,.$$

In practice, one never tests independence of random variables using the above multiplication law.

Proposition 3.4. *Let $X_1, \dots, X_N$ be N random variables. The following are equivalent:*

(1) The random variables $X_1, \dots, X_N$ are independent.

(2) For every $\alpha_1, \dots, \alpha_N \in \mathbb{R}$, we have

$$\boldsymbol{P}\Big(\bigcap_{j=1}^{N} \{X_j \leqslant \alpha_j\}\Big) = \prod_{j=1}^{N} \boldsymbol{P}(X_j \leqslant \alpha_j)$$

(3) For every collection of bounded continuous functions $f_1, \dots, f_N$ we have

$$\boldsymbol{E}\Big[\prod_{j=1}^{N} f_j(X_j)\Big] = \prod_{j=1}^{N} \boldsymbol{E} f_j(X_j)\,.$$

(4) For every $\xi_1, \dots, \xi_N \in \mathbb{R}$ we have

$$\boldsymbol{E}\exp\Big(i\sum_{j=1}^{N} \xi_j X_j\Big) = \prod_{j=1}^{N} \boldsymbol{E}\exp(i\xi_j X_j)\,, \qquad \textit{where } i = \sqrt{-1}\,.$$

Remark 3.5. It is instructive to explicitly check each of these implications when $N = 2$ and X_1, X_2 are simple random variables.

Remark 3.6. The intuition behind the above result is as follows: Since the events $\{X_j \leqslant \alpha_j\}$ generate $\sigma(X_j)$, we expect the first two properties to be equivalent. Since $\mathbf{1}_{(-\infty,\alpha_j]}$ can be well approximated by continuous functions, we expect equivalence of the second and third properties. The last property is a bit more subtle: Since $\exp(a + b) = \exp(a)\exp(b)$, the third clearly implies the last property. The converse holds because of "completeness of the complex exponentials" or Fourier inversion, and again a through discussion of this will lead us too far astray.

Remark 3.7. The third implication above implies that independent random variables are uncorrelated. Namely, if X, Y are independent random variables, then

$$\boldsymbol{E}(XY) = (\boldsymbol{E}X)(\boldsymbol{E}Y)\,. \tag{3.1}$$

The converse, is of course false. Namely if (3.1) holds, there is no reason we should have

$$\boldsymbol{E} f(X) g(Y) = \boldsymbol{E} f(X) \boldsymbol{E} g(Y)\,,$$

for *every* bounded continuous pair of functions f, g as required by the third part in Proposition 3.4. However, if (X, Y) is *jointly normal* and X, Y are uncorrelated, then the *normal correlation theorem* guarantees that X, Y are independent.

Remark 3.8. If moment generating functions of the random variables are defined in an interval around 0, then one can test independence using real exponentials instead of the complex exponentials used in the last condition in Proposition 3.4. Explicitly, in this case $X_1, \dots, X_N$ are independent if and only if for every $t_1, \dots, t_N$ in some small interval containing 0 we have

$$\boldsymbol{E}\exp\Big(\sum_{j=1}^{N} t_j X_j\Big) = \prod_{j=1}^{N} \boldsymbol{E}\exp(t_j X_j)\,.$$

Example 3.9 (Covariance of Brownian motion). The independence of increments allows us to compute covariances of Brownian motion easily. Suppose W is a standard Brownian motion, and $s < t$. Then we know $W_s \sim N(0, s)$, $W_t - W_s \sim N(0, t - s)$ and is independent of W_s. Consequently $(W_s, W_t - W_s)$ is jointly normal with mean 0 and covariance matrix $\left(\begin{smallmatrix} s & 0 \\ 0 & t-s \end{smallmatrix}\right)$. This implies that (W_s, W_t) is a jointly normal random variable. Moreover we can compute the covariance by

$$\boldsymbol{E} W_s W_t = \boldsymbol{E} W_s (W_t - W_s) + \boldsymbol{E} W_s^2 = s\,.$$

In general if you don't assume $s < t$, the above immediately implies $\boldsymbol{E} W_s W_t = s \wedge t$.

4. Conditional probability

Our next goal is to understand *conditional probability*, and we do it directly here to help understanding. In the next section we will construct *conditional expectations* independently, and the reader may choose to skip this section.

Suppose you have an incomplete deck of cards which has 10 red cards, and 20 black cards. Suppose 5 of the red cards are *high cards* (i.e. ace, king, queen, jack or 10), and only 4 of the black cards are high. If a card is chosen at random, the *conditional probability* of it being high given that it is red is 1/2, and the *conditional probability* of it being high given that it is black is 1/5. Our aim is to encode both these facts into a single entity.

We do this as follows. Let R, B denote the set of all red and black cards respectively, and H denote the set of all high cards. A σ-algebra encompassing all the above information is exactly

$$\mathcal{G} \stackrel{\text{def}}{=} \big\{\emptyset, R, B, H, H^c, R \cap H, B \cap H, R \cap H^c, B \cap H^c, \\ (R \cap H) \cup (B \cap H^c), (R \cap H^c) \cup (B \cap H), \Omega\big\}$$

and you can explicitly compute the probabilities of each of the above events. A σ-algebra encompassing only the color of cards is exactly

$$\mathcal{G} \stackrel{\text{def}}{=} \{\emptyset, R, B, \Omega\}\,.$$

Now we define the *conditional probability* of a card being high given the color to be the **random variable**

$$\boldsymbol{P}(H \mid \mathcal{C}) \stackrel{\text{def}}{=} \boldsymbol{P}(H \mid R)\mathbf{1}_R + \boldsymbol{P}(H \mid B)\mathbf{1}_B = \frac{1}{2}\mathbf{1}_R + \frac{1}{5}\mathbf{1}_B\,.$$

To emphasize:

1. What is given is the σ-algebra $\mathcal{C}$, and not just an event.
2. The conditional probability is now a *$\mathcal{C}$-measurable random variable* and not a number.

To see how this relates to $\boldsymbol{P}(H \mid R)$ and $\boldsymbol{P}(H \mid B)$ we observe

$$\int_R \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} \stackrel{\text{def}}{=} \boldsymbol{E}\big(\mathbf{1}_R \boldsymbol{P}(H \mid \mathcal{C})\big) = \boldsymbol{P}(H \mid R)\, \boldsymbol{P}(R)\,.$$

The same calculation also works for B, and so we have

$$\boldsymbol{P}(H \mid R) = \frac{1}{\boldsymbol{P}(R)} \int_R \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} \quad \text{and} \quad \boldsymbol{P}(H \mid B) = \frac{1}{\boldsymbol{P}(B)} \int_B \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P}\,.$$

Our aim is now to generalize this to a non-discrete scenario. The problem with the above identities is that if either R or B had probability 0, then the above would become meaningless. However, clearing out denominators yields

$$\int_R \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} = \boldsymbol{P}(H \cap R) \quad \text{and} \quad \int_B \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} = \boldsymbol{P}(H \cap B)\,.$$

This suggests that the defining property of $\boldsymbol{P}(H \mid \mathcal{C})$ should be the identity

$$\int_C \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} = \boldsymbol{P}(H \cap C) \tag{4.1}$$

for every event $C \in \mathcal{C}$. Note $\mathcal{C} = \{\emptyset, R, B, \Omega\}$ and we have only checked (4.1) for $C = R$ and $C = B$. However, for $C = \emptyset$ and $C = \Omega$, (4.1) is immediate.

Definition 4.1. Let $(\Omega, \mathcal{G}, \boldsymbol{P})$ be a probability space, and $\mathcal{F} \subseteq \mathcal{G}$ be a σ-algebra. Given $A \in \mathcal{G}$, we define the conditional probability of A given $\mathcal{F}$, denoted by $\boldsymbol{P}(A|\mathcal{F})$ to be an $\mathcal{F}$-measurable random variable that satisfies

$$\int_F \boldsymbol{P}(H \mid \mathcal{F})\, d\boldsymbol{P} = \boldsymbol{P}(H \cap F) \qquad \text{for every } F \in \mathcal{F}. \tag{4.2}$$

Remark 4.2. Showing existence (and uniqueness) of the conditional probability isn't easy, and relies on the *Radon-Nikodym theorem*, which is beyond the scope of this course.

Remark 4.3. It is crucial to require that $\boldsymbol{P}(H|\mathcal{F})$ is measurable with respect to $\mathcal{F}$. Without this requirement we could simply choose $\boldsymbol{P}(H|\mathcal{F}) = \boldsymbol{1}_H$ and (4.2) would be satisfied. However, note that if $H \in \mathcal{F}$, then the function $\boldsymbol{1}_F$ is $\mathcal{F}$-measurable, and in this case $\boldsymbol{P}(H \mid \mathcal{F}) = \boldsymbol{1}_F$.

Remark 4.4. In general we can only expect (4.2) to hold for all events in $\mathcal{F}$, and it need not hold for events in $\mathcal{G}$! Indeed, in the example above we see that

$$\int_H \boldsymbol{P}(H \mid \mathcal{C})\, d\boldsymbol{P} = \frac{1}{2}\boldsymbol{P}(R \cap H) + \frac{1}{5}\boldsymbol{P}(B \cap H) = \frac{1}{2} \cdot \frac{5}{30} + \frac{1}{5} \cdot \frac{4}{30} = \frac{11}{100}$$

but

$$\boldsymbol{P}(H \cap H) = \boldsymbol{P}(H) = \frac{3}{10} \neq \frac{11}{100}\,.$$

Remark 4.5. One situation where you can compute $\boldsymbol{P}(A|\mathcal{F})$ explicitly is when $\mathcal{F} = \sigma(\{F_i\})$ where $\{F_i\}$ is a pairwise disjoint collection of events whose union is all of Ω and $\boldsymbol{P}(F_i) > 0$ for all i. In this case

$$\boldsymbol{P}(A \mid \mathcal{F}) = \sum_i \frac{\boldsymbol{P}(A \cap F_i)}{\boldsymbol{P}(F_i)} \boldsymbol{1}_{F_i}\,.$$

5. Conditional expectation.

Conditional expectation arises when you have a random variable X, and want to *best approximate* it using only a (strict) subset of events. Precisely, suppose $\mathcal{F} \subseteq \mathcal{G}$ is a *σ-sub-algebra* of $\mathcal{G}$. That is, $\mathcal{F}$ is a σ-algebra, and every event in $\mathcal{F}$ is also an event in $\mathcal{G}$. Now to best approximate a ($\mathcal{G}$-measurable) random variable X using only events in $\mathcal{F}$, one would like to find an $\mathcal{F}$ measurable random variable Z that minimizes

$$\boldsymbol{E}|X - Z|^2\,.$$

The minimizer is known as the *conditional expectation of X given $\mathcal{F}$*, and denoted by $\boldsymbol{E}(X \mid \mathcal{F})$. That is,

$$\boldsymbol{E}(X \mid \mathcal{F}) \overset{\text{def}}{=} \arg\min\bigl\{\boldsymbol{E}|X - Z|^2 \bigm| Z \text{ is a } \mathcal{G}\text{-measurable random variable}\bigr\}\,. \tag{5.1}$$

While the above provides good intuition to the notion of conditional expectation, it is not as convenient to work with mathematically. For instance, the above requires $\boldsymbol{E}X^2 < \infty$, and we will often require conditional expectations of random variables that do not have this property.

To motivate the other definition of conditional expectation, we use the following example. Consider an incomplete deck of cards which has 10 red cards, of which 5 are high, and 20 black cards, of which 4 are high. Let X be the outcome of a game played through a dealer who pays you \$1 when a high card is drawn, and charges you \$1 otherwise. However, you are standing too far away from the dealer to tell whether the card drawn was high or not. You can only tell *the color*, and *whether or not you won.*

After playing this game often the only information you can deduce is that your expected return is 0 when a red card is drawn and $-3/5$ when a black card is drawn. That is, you approximate the game outcome X by the random variable

$$Y \overset{\text{def}}{=} 0\boldsymbol{1}_R - \frac{3}{5}\boldsymbol{1}_B\,,$$

where, as before R, B denote the set of all red and black cards respectively.

Note that the events you can deduce information about by playing this game (through the dealer) are exactly elements of the σ-algebra $\mathcal{C} = \{\emptyset, R, B, \Omega\}$. By construction, that your approximation Y is $\mathcal{C}$-measurable, and it is easy to verify that

$$Y = \arg\min\{\boldsymbol{E}(X - Z)^2 \mid Z \text{ is a } \mathcal{C}\text{-measurable random variable}\}\,. \tag{5.2}$$

That is $Y = \boldsymbol{E}(X \mid C)$ according to the definition (5.1). In this case, we can also verify that Y has the same averages as X on all elements of $\mathcal{C}$. That is, for every $C \in \mathcal{C}$, we have[4]

$$\int_C Y\, d\boldsymbol{P} = \int_C X\, d\boldsymbol{P}\,. \tag{5.3}$$

It turns out that in general, one can show abstractly that any $\mathcal{C}$ measurable random variable that satisfies (5.3), must in fact also be the minimizer in (5.2). We will thus use (5.3) to define conditional expectation.

Definition 5.1. Let X be a $\mathcal{G}$-measurable random variable, and $\mathcal{F} \subseteq \mathcal{G}$ be a σ-sub-algebra. We define $\boldsymbol{E}(X \mid \mathcal{F})$, the *conditional expectation of X given $\mathcal{F}$* to be a *random variable* such that:

(1) $\boldsymbol{E}(X \mid \mathcal{F})$ is $\mathcal{F}$-measurable.
(2) For every $F \in \mathcal{F}$, we have the *partial averaging* identity:

$$\int_F \boldsymbol{E}(X \mid \mathcal{F})\, d\boldsymbol{P} = \int_F X\, d\boldsymbol{P}. \tag{5.4}$$

[4] Recall $\int_C Y\, d\boldsymbol{P}$ is simply $\boldsymbol{E}(\boldsymbol{1}_C Y)$. That is $\int_C Y\, d\boldsymbol{P}$ is the expectation of the random variable which is Y on the event C, and 0 otherwise.

REMARK 5.2. We can only expect (5.4) to hold for all events $F \in \mathcal{F}$. In general (5.4) *will not* hold for events $G \in \mathcal{G} - \mathcal{F}$.

REMARK 5.3. An equivalent way of phrasing (5.4) is to require

$$\boldsymbol{E}(XY) = \boldsymbol{E}\big(\boldsymbol{E}(X \mid \mathcal{F})Y\big) \tag{5.5}$$

for every $\mathcal{F}$ measurable random variable Y. As before, we can only expect (5.5) to hold when Y is $\mathcal{F}$ measurable. In general (5.5) *will not* hold when Y is not $\mathcal{F}$ measurable.

REMARK 5.4. Choosing $F = \Omega$ we see $\boldsymbol{E}\boldsymbol{E}(X \mid \mathcal{F}) = \boldsymbol{E}X$.

REMARK 5.5. More concretely, suppose Y is another random variable and $\mathcal{F} = \sigma(Y)$. Then it turns out that one can find a special (non-random) function g such that $\boldsymbol{E}(X \mid \mathcal{F}) = g(Y)$. Moreover, the function g is characterized by the property that

$$\boldsymbol{E}\big(f(Y)X\big) = \boldsymbol{E}\big(f(Y)g(Y)\big).$$

for *every* bounded continuous function f.

REMARK 5.6. Under mild integrability assumptions one can show that conditional expectations exist. This requires the *Radon-Nikodym* theorem and goes beyond the scope of this course. If, however, $\mathcal{F} = \sigma(\{F_i\})$ where $\{F_i\}$ is a pairwise disjoint collection of events whose union is all of Ω and $\boldsymbol{P}(F_i) > 0$ for all i, then

$$\boldsymbol{E}(X \mid \mathcal{F}) = \sum_{i=1}^{\infty} \frac{\boldsymbol{1}_{F_i}}{\boldsymbol{P}(F_i)} \int_{F_i} X \, d\boldsymbol{P}.$$

REMARK 5.7. Once existence is established it is easy to see that conditional expectations are unique. Namely, if Y is any $\mathcal{F}$-measurable random variable that satisfies

$$\int_F Y \, d\boldsymbol{P} = \int_F X \, d\boldsymbol{P} \quad \text{for every } F \in \mathcal{F},$$

then $Y = \boldsymbol{E}(X \mid F)$. Often, when computing the conditional expectation, we will "guess" what it is, and verify our guess by checking measurablity and the above partial averaging identity.

PROPOSITION 5.8. *If X is $\mathcal{F}$-measurable, then $\boldsymbol{E}(X \mid \mathcal{F}) = X$. On the other hand, if X is independent[5] of $\mathcal{F}$ then $\boldsymbol{E}(X \mid \mathcal{F}) = \boldsymbol{E}X$.*

PROOF. If X is $\mathcal{F}$-measurable, then clearly the random variable X is both $\mathcal{F}$-measurable and satisfies the partial averaging identity. Thus by uniqueness, we must have $\boldsymbol{E}(X \mid \mathcal{F}) = X$.

Now consider the case when X is independent of $\mathcal{F}$. Suppose first $X = \sum a_i \boldsymbol{1}_{A_i}$ for finitely many sets $A_i \in \mathcal{G}$. Then for any $F \in \mathcal{F}$,

$$\int_F X \, d\boldsymbol{P} = \sum a_i \boldsymbol{P}(A_i \cap F) = \boldsymbol{P}(F) \sum a_i \boldsymbol{P}(A_i) = \boldsymbol{P}(F)\boldsymbol{E}X = \int_F \boldsymbol{E}X \, d\boldsymbol{P}.$$

Thus the constant random variable $\boldsymbol{E}X$ is clearly $\mathcal{F}$-measurable and satisfies the partial averaging identity. This forces $\boldsymbol{E}(X \mid \mathcal{F}) = \boldsymbol{E}X$. The general case when X is not simple follows by approximation. □

The above fact has a generalization that is tremendously useful when computing conditional expectations. Intuitively, the general principle is to *average* quantities that are independent of $\mathcal{F}$, and *leave unchanged* quantities that are $\mathcal{F}$ measurable. This is known as the independence lemma.

LEMMA 5.9 (Independence Lemma). *Suppose X, Y are two random variables such that X is independent of $\mathcal{F}$ and Y is $\mathcal{F}$-measurable. Then if $f = f(x, y)$ is any function of two variables we have*

$$\boldsymbol{E}\big(f(X, Y) \bigm| \mathcal{F}\big) = g(Y),$$

where $g = g(y)$ is the function[6] defined by

$$g(y) \stackrel{\text{def}}{=} \boldsymbol{E}f(X, y).$$

REMARK. If p_X is the probability density function of X, then the above says

$$\boldsymbol{E}\big(f(X, Y) \bigm| \mathcal{F}\big) = \int_{\mathbb{R}} f(x, Y)\, p_X(x) \, dx.$$

Indicating the ω dependence explicitly for clarity, the above says

$$\boldsymbol{E}\big(f(X, Y) \bigm| \mathcal{F}\big)(\omega) = \int_{\mathbb{R}} f(x, Y(\omega))\, p_X(x) \, dx.$$

REMARK 5.10. Note we defined and motivated conditional expectations and conditional probabilities independently. They are however intrinsically related: Indeed, $\boldsymbol{E}(\boldsymbol{1}_A \mid \mathcal{F}) = \boldsymbol{P}(A \mid \mathcal{F})$, and this can be checked directly from the definition.

As we will see shortly, computing conditional expectations will be a very important part of pricing securities. Most of the time, all that is required to compute conditional expectations are the following properties.

PROPOSITION 5.11. *Conditional expectations satisfy the following properties.*

(1) (Linearity) *If X, Y are random variables, and $\alpha \in \mathbb{R}$ then*

$$\boldsymbol{E}(X + \alpha Y \mid \mathcal{F}) = \boldsymbol{E}(X \mid \mathcal{F}) + \alpha \boldsymbol{E}(Y \mid \mathcal{F}).$$

(2) (Positivity) *If $X \leqslant Y$, then $\boldsymbol{E}(X \mid \mathcal{F}) \leqslant \boldsymbol{E}(Y \mid \mathcal{F})$ (almost surely).*

(3) If X is $\mathcal{F}$ measurable and Y is an arbitrary (not necessarily $\mathcal{F}$-measurable) random variable then (almost surely)

$$\boldsymbol{E}(XY \mid \mathcal{F}) = X\boldsymbol{E}(Y \mid \mathcal{F}).$$

(4) (Tower property) *If $\mathcal{E} \subseteq \mathcal{F} \subseteq \mathcal{G}$ are σ-algebras, then (almost surely)*

$$\boldsymbol{E}(X \mid \mathcal{E}) = \boldsymbol{E}\Big(\boldsymbol{E}(X \mid \mathcal{F}) \Bigm| \mathcal{E}\Big).$$

[5] We say a random variable X is independent of σ-algebra $\mathcal{F}$ if for every $A \in \sigma(X)$ and $B \in \mathcal{F}$ the events A and B are independent.

[6] To clarify, we are defining a *non-random* function $g = g(y)$ here when $y \in \mathbb{R}$ is any real number. Then, once we compute g, we substitute in $y = Y(= Y(\omega))$, where Y is the given random variable.

Proof. The first property follows immediately from linearity. For the second property, set $Z = Y - X$ and observe

$$\int_{\boldsymbol{E}(Z|\mathcal{F})} \boldsymbol{E}(Z \mid \mathcal{F}) \, d\boldsymbol{P} = \int_{\boldsymbol{E}(Z|\mathcal{F})} Z \, d\boldsymbol{P} \geqslant 0 \,,$$

which can only happen if $\boldsymbol{P}(\boldsymbol{E}(Z \mid \mathcal{F}) < 0) = 0$. The third property is easily checked for simple random variables, and follows in general by approximating. The tower property follows immediately from the definition. □

As an illustration, of how the above properties come in handy, we show how they can be used to deduce (5.5).

Proposition 5.12. *If $\mathcal{F} \subseteq G$ is a σ-sub-algebra, X is a $\mathcal{G}$-measurable random variable, and Y is an $\mathcal{F}$-measurable random variable, then*

$$\boldsymbol{E}(XY) = \boldsymbol{E}\big(\boldsymbol{E}(X \bigm| \mathcal{F})Y\big)$$

Proof. Using Remark 5.4 and then Proposition 5.11 part (3), we see

$$\boldsymbol{E}(XY) = \boldsymbol{E}\big(\boldsymbol{E}(XY \mid \mathcal{F})\big) = \boldsymbol{E}\big(Y \boldsymbol{E}(X \mid \mathcal{F})\big) \,,$$

as desired. □

Finally we conclude this section by showing that the conditional expectation of a random variable according to Definition 5.1 is precisely the minimizer as in (5.1).

Proposition 5.13. *Let X be a square integrable $\mathcal{G}$-measurable random variable, and $\mathcal{F} \subseteq \mathcal{G}$ be a σ-sub-algebra of $\mathcal{G}$. Then amongst all $\mathcal{F}$-measurable random variables Z, the one that minimizes*

$$\boldsymbol{E}(X - Z)^2$$

is precisely $Z = \boldsymbol{E}(X \mid \mathcal{F})$.

Proof. Since $\boldsymbol{E}(X \mid \mathcal{F})$ is known to be an $\mathcal{F}$-measurable random variable, we only need to show that for any (other) $\mathcal{F}$-measurable random variable $\mathbb{Z}$ we have

$$\boldsymbol{E}(X - Z)^2 \geqslant \boldsymbol{E}\big((X - \boldsymbol{E}(X \bigm| \mathcal{F}))^2\big) \,.$$

To see this, note

$$\begin{aligned} \boldsymbol{E}(X - Z)^2 &= \boldsymbol{E}(X - \boldsymbol{E}(X \mid \mathcal{F}) + \boldsymbol{E}(X \mid \mathcal{F}) - Z)^2 \\ &= \boldsymbol{E}(X - \boldsymbol{E}(X \mid \mathcal{F}))^2 + \boldsymbol{E}((X \mid \mathcal{F}) - Z)^2 \\ &\qquad + 2\boldsymbol{E}\Big(\underbrace{(X - \boldsymbol{E}(X \mid \mathcal{F}))}_{I}\underbrace{(\boldsymbol{E}(X \mid \mathcal{F}) - Z)}_{II}\Big) \end{aligned}$$

Since term II is $\mathcal{F}$ measurable, we can use (5.5) to replace X with $\boldsymbol{E}(X \mid F)$ in term I. This yields

$$\begin{aligned} \boldsymbol{E}(X - Z)^2 &= \boldsymbol{E}(X - \boldsymbol{E}(X \mid \mathcal{F}))^2 + \boldsymbol{E}((X \mid \mathcal{F}) - Z)^2 \\ &\qquad + 2\boldsymbol{E}\Big((\boldsymbol{E}(X \mid \mathcal{F}) - \boldsymbol{E}(X \mid \mathcal{F}))(\boldsymbol{E}(X \mid \mathcal{F}) - Z)\Big) \\ &= \boldsymbol{E}(X - \boldsymbol{E}(X \mid \mathcal{F}))^2 + \boldsymbol{E}((X \mid \mathcal{F}) - Z)^2 \geqslant \boldsymbol{E}(X - \boldsymbol{E}(X \mid \mathcal{F}))^2 \,, \end{aligned}$$

as desired. □

6. The Martingale Property

A martingale is "fair game". Suppose you are playing a game and $M(t)$ is your cash stockpile at time t. As time progresses, you learn more and more information about the game. For instance, in blackjack getting a high card benefits the player more than the dealer, and a common card counting strategy is to have a "spotter" betting the minimum while counting the high cards. When the odds of getting a high card are favorable enough, the player will signal a "big player" who joins the table and makes large bets, as long as the high card count is favorable. Variants of this strategy have been shown to give the player up to a 2% edge over the house.

If a game is a martingale, then this extra information you have acquired *can not* help you going forward. That is, if you signal your "big player" at any point, you will not affect your expected return.

Mathematically this translates to saying that the *conditional expectation* of your stockpile at a later time given your present accumulated knowledge, is exactly the present value of your stockpile. Our aim in this section is to make this precise.

6.1. Adapted processes and filtrations. Let X be any stochastic process (for example Brownian motion). For any $t > 0$, we've seen before that $\sigma(X(t))$ represents the information you obtain by observing $X(t)$. Accumulating this over time gives us the *filtration*. To introduce this concept, we first need the notion of a σ algebra generated by a family of sets.

Definition 6.1. Given a collection of sets A_α, where α belongs to some (possibly infinite) index set $\mathcal{A}$, we define $\sigma(\{A_\alpha\})$ to be the *smallest* σ-algebra that contains each of the sets A_α.

That is, if $\mathcal{G} = \sigma(\{A_\alpha\})$, then we must have each $A_\alpha \in \mathcal{G}$. Since $\mathcal{G}$ is a σ-algebra, all sets you can obtain from these by taking complements, countable unions and countable intersections intersections must also belong to $\mathcal{G}$.[7] The fact that $\mathcal{G}$ is the smallest σ-algebra containing each A_α also means that if $\mathcal{G}'$ is any other σ-algebra that contains each A_α, then $\mathcal{G} \subseteq \mathcal{G}'$.

Remark 6.2. The smallest σ-algebra under which X is a random variable (under which X is measurable) is exactly $\sigma(X)$. It turns out that $\sigma(X) = X^{-1}(\mathcal{B}) = \{X \in B | B \in \mathcal{B}\}$, where $\mathcal{B}$ is the *Borel* σ-algebra on $\mathbb{R}$. Here $\mathcal{B}$ is the *Borel* σ-algebra, defined to be the σ-algebra on $\mathbb{R}$ generated by all open intervals.

Definition 6.3. Given a stochastic process X, the *filtration generated by X* is the family of σ-algebras $\{\mathcal{F}_t^X \mid t \geqslant 0\}$ where

$$\mathcal{F}_t^X \overset{\text{def}}{=} \sigma\Big(\bigcup_{s \leqslant t} \sigma(X_s)\Big) \,.$$

[7] Usually $\mathcal{G}$ contains *much more* than all countable unions, intersections and complements of the A_α's. You might think you could keep including all sets you generate using countable unions and complements and arrive at all of $\mathcal{G}$. It turns out that to make this work, you will usually have to do this *uncountably* many times!

This won't be too important within the scope of these notes. However, if you read a rigorous treatment and find the authors using some fancy trick (using Dynkin systems or monotone classes) instead of a naive countable unions argument, then the above is the reason why.

That is, $\mathcal{F}_t^X$ is all events that can be observed using only the random variables X_s when $s \leqslant t$. Clearly each $\mathcal{F}_t^X$ is a σ-algebra, and if $s \leqslant t$, $\mathcal{F}_s^X \subseteq \mathcal{F}_t^X$. A family of σ-algebras with this property is called a *filtration.*

DEFINITION 6.4. A *filtration* is a family of σ-algebras $\{\mathcal{F}_t \mid t \geqslant 0\}$ such that whenever $s \leqslant t$, we have $\mathcal{F}_s \subseteq \mathcal{F}_t$.

In our case, the filtration we work with will most often be the *Brownian filtration*, i.e. the filtration generated by Brownian motion. However, one can (and often needs to) consider more general filtrations. In this case the intuition we use is that the σ-algebra $\mathcal{F}_t$ represents the information accumulated up to time t (i.e. all events whose probabilities can be deduced up to time t). When given a filtration, it is important that all stochastic processes we construct respect the flow of information, and do not look into the future. This is of course natural: trading / pricing strategies can not rely on the price at a later time, and gambling strategies do not know the outcome of the next hand. Mathematically this property is called *adapted*, and is defined as follows.

DEFINITION 6.5. A stochastic process X is said to be *adapted* to a filtration $\{\mathcal{F}_t \mid t \geqslant 0\}$ if for every t the random variable $X(t)$ is $\mathcal{F}_t$ measurable (i.e. $\{X(t) \leqslant \alpha\} \in \mathcal{F}_t$ for every $\alpha \in \mathbb{R}$, $t \geqslant 0$).

Clearly a process X is adapted with respect to the filtration it generates $\{\mathcal{F}_t^X\}$.

6.2. Martingales. Recall, a martingale is a "fair game". Using conditional expectations, we can now define this precisely.

DEFINITION 6.6. A stochastic process M is a martingale with respect to a filtration $\{\mathcal{F}_t\}$ if:

(1) M is adapted to the filtration $\{\mathcal{F}_t\}$.
(2) For any $s < t$ we have $\boldsymbol{E}(M(t) \mid \mathcal{F}_s) = M(s)$, almost surely.

REMARK 6.7. A *sub-martingale* is an adapted process M for which we have $\boldsymbol{E}(M(t) \mid \mathcal{F}_s) \geqslant M(s)$, and a *super-martingale* if $\boldsymbol{E}(M(t) \mid \mathcal{F}_s) \leqslant M(s)$. Thus $\boldsymbol{E}M(t)$ is an increasing function of time if M is a sub-martingale, constant in time if M is a martingale, and a decreasing function of time if M is a super-martingale.

REMARK 6.8. It is crucial to specify the filtration when talking about martingales, as it is certainly possible that a process is a martingale with respect to one filtration but not with respected to another. For our purposes the filtration will almost always be the *Brownian filtration* (i.e. the filtration generated by Brownian motion).

EXAMPLE 6.9. Let $\{\mathcal{F}_t\}$ be a filtration, $\mathcal{F}_\infty = \sigma(\cup_{t \geqslant 0} \mathcal{F}_t)$, and X be any $\mathcal{F}_\infty$-measurable random variable. The process $M(t) \stackrel{\text{def}}{=} \boldsymbol{E}(X_\infty \mid \mathcal{F}_t)$ is a martingale with respect to the filtration $\{\mathcal{F}_t\}$.

6.3. The martingale property of Brownian motion. In discrete time a random walk is a martingale, so it is natural to expect that in continuous time Brownian motion is a martingale as well.

THEOREM 6.10. *Let W be a Brownian motion, $\mathcal{F}_t = \mathcal{F}_t^W$ be the Brownian filtration. Brownian motion is a martingale with respect to this filtration.*

PROOF. By independence of increments, $W(t) - W(s)$ is certainly independent of $W(r)$ for any $r \leqslant s$. Since $\mathcal{F}_s = \sigma(\cup_{r \leqslant s} \sigma(W(r)))$ we expect that $W(t) - W(s)$ is independent of $\mathcal{F}_s$. Consequently

$$\boldsymbol{E}(W(t) \mid \mathcal{F}_s) = \boldsymbol{E}(W(t) - W(s) \mid \mathcal{F}_s) + \boldsymbol{E}(W(s) \mid \mathcal{F}_s) = 0 + W(s) = W(s)\,. \qquad \square$$

THEOREM 6.11. *Let W be a standard Brownian motion (i.e. a Brownian motion normalized so that $W(0) = 0$ and $\operatorname{Var}(W(t)) = t$). For any $C_b^{1,2}$ function*[8] *$f = f(t,x)$ the process*

$$M(t) \stackrel{\text{def}}{=} f(t, W(t)) - \int_0^t \Bigl(\partial_t f(s, W(s)) + \frac{1}{2}\partial_x^2 f(s, W(s))\Bigr)\, ds$$

is a martingale (with respect to the Brownian filtration).

PROOF. This is an extremely useful fact about Brownian motion follows quickly from the Itô formula, which we will discuss later. However, at this stage, we can provide a simple, elegant and instructive proof as follows.

Adaptedness of M is easily checked. To compute $\boldsymbol{E}(M(t) \mid \mathcal{F}_r)$ we first observe

$$\boldsymbol{E}\bigl(f(t, W(t)) \bigm| \mathcal{F}_r\bigr) = \boldsymbol{E}\bigl(f(t, [W(t) - W(r)] + W(r)) \bigm| \mathcal{F}_r\bigr).$$

Since $W(t) - W(r) \sim N(0, t-r)$ and is independent of $\mathcal{F}_r$, the above conditional expectation can be computed by

$$\boldsymbol{E}\bigl(f(t, [W(t) - W(r)] + W(r)) \bigm| \mathcal{F}_r\bigr) = \int_{\mathbb{R}} f(t, y + W(r)) G(t-r, y)\, dy\,,$$

where

$$G(\tau, y) = \frac{1}{\sqrt{2\pi\tau}} \exp\Bigl(\frac{-y^2}{2\tau}\Bigr)$$

is the density of $W(t) - W(r)$.

Similarly

$$\begin{aligned} &\boldsymbol{E}\Bigl(\int_0^t \bigl(\partial_t f(s, W(s)) + \frac{1}{2}\partial_x^2 f(s, W(s))\bigr)\, ds \Bigm| \mathcal{F}_r\Bigr) \\ &\quad = \int_0^r \bigl(\partial_t f(s, W(s)) + \frac{1}{2}\partial_x^2 f(s, W(s))\bigr)\, ds \\ &\qquad + \int_r^t \int_{\mathbb{R}} \bigl(\partial_t f(s, y + W(r)) + \frac{1}{2}\partial_x^2 f(s, y + W(r))\bigr) G(s-r, y)\, dy\, ds \end{aligned}$$

Hence

$$\boldsymbol{E}(M(t) \mid \mathcal{F}_r) - M(r) = \int_{\mathbb{R}} f(t, y + W(r)) G(t-r, y)\, dy$$

[8] Recall a function $f = f(t,x)$ is said to be $C^{1,2}$ if it is C^1 in t (i.e. differentiable with respect to t and $\partial_t f$ is continuous), and C^2 in x (i.e. twice differentiable with respect to x and $\partial_x f$, $\partial_x^2 f$ are both continuous). The space $C_b^{1,2}$ refers to all $C^{1,2}$ functions f for which and f, $\partial_t f$, $\partial_x f$, $\partial_x^2 f$ are all bounded functions.

$$
\begin{aligned}
&- \int_r^t \int_{\mathbb{R}} \Big(\partial_t f(s, y + W(r)) + \frac{1}{2}\partial_x^2 f(s, y + W(r))\Big) G(s-r, y)\, ds \\
&- f(r, W(r))\,.
\end{aligned}
$$

We claim that the right hand side of the above vanishes. In fact, we claim the (deterministic) identity

$$
\begin{aligned}
f(r,x) = &\int_{\mathbb{R}} f(t, y+x) G(t-r, y)\, dy \\
&\qquad - \int_r^t \int_{\mathbb{R}} \Big(\partial_t f(s, y+x) + \frac{1}{2}\partial_x^2 f(s, y+x)\Big)\Big) G(s-r, y)\, dy\, ds
\end{aligned}
$$

holds for any function f and $x \in \mathbb{R}$. For those readers who are familiar with PDEs, this is simply the Duhamel's principle for the heat equation. If you're unfamiliar with this, the above identity can be easily checked using the fact that $\partial_\tau G = \frac{1}{2}\partial_y^2 G$ and integrating the first integral by parts. We leave this calculation to the reader. □

6.4. Stopping Times. For this section we assume that a filtration $\{\mathcal{F}_t\}$ is given to us, and fixed. When we refer to process being adapted (or martingales), we implicitly mean they are adapted (or martingales) with respect to this filtration.

Consider a game (played in continuous time) where you have the option to walk away at any time. Let τ be the *random* time you decide to stop playing and walk away. In order to respect the flow of information, you need to be able to decide weather you have stopped using only information up to the present. At time t, event $\{\tau \leqslant t\}$ is exactly when you have stopped and walked away. Thus, to respect the flow of information, we need to ensure $\{\tau \leqslant t\} \in \mathcal{F}_t$.

Definition 6.12. A stopping time is a function $\tau\colon \Omega \to [0, \infty)$ such that for every $t \geqslant 0$ the event $\{\tau \leqslant t\} \in \mathcal{F}_t$.

A standard example of a stopping time is *hitting times*. Say you decide to liquidate your position once the value of your portfolio reaches a certain threshold. The time at which you liquidate is a hitting time, and under mild assumptions on the filtration, will always be a stopping time.

Proposition 6.13. *Let X be an adapted continuous process, $\alpha \in \mathbb{R}$ and τ be the first time X hits α (i.e. $\tau = \inf\{t \geqslant 0 \mid X(t) = \alpha\}$). Then τ is a stopping time (if the filtration is right continuous).*

Theorem 6.14 (Doob's optional sampling theorem). *If M is a martingale and τ is a bounded stopping time. Then the stopped process $M^\tau(t) \stackrel{\text{def}}{=} M(\tau \wedge t)$ is also a martingale. Consequently, $\boldsymbol{E}M(\tau) = \boldsymbol{E}M(\tau \wedge t) = \boldsymbol{E}M(0) = \boldsymbol{E}M(t)$ for all $t \geqslant 0$.*

Remark 6.15. If instead of assuming τ is bounded, we assume M^τ is bounded the above result is still true.

The proof goes beyond the scope of these notes, and can be found in any standard reference. What this means is that if you're playing a fair game, then you can not hope to improve your odds by "quitting when you're ahead". Any rule by which you decide to stop, must be a stopping time and the above result guarantees that stopping a martingale still yields a martingale.

Remark 6.16. Let W be a standard Brownian motion, τ be the first hitting time of W to 1. Then $\boldsymbol{E}W(\tau) = 1 \neq 0 = \boldsymbol{E}W(t)$. This is one situation where the optional sampling theorem doesn't apply (in fact, $\boldsymbol{E}\tau = \infty$, and W^τ is unbounded).

This example corresponds to the gambling strategy of walking away when you make your "million". The reason it's not a sure bet is because the time taken to achieve your winnings is finite almost surely, but very long (since $\boldsymbol{E}\tau = \infty$). In the mean time you might have incurred financial ruin and expended your entire fortune.

Suppose the price of a security you're invested in fluctuates like a martingale (say for instance Brownian motion). This is of course unrealistic, since Brownian motion can also become negative; but lets use this as a first example. You decide you're going to liquidate your position and walk away when either you're bankrupt, or you make your first million. What are your expected winnings? This can be computed using the optional sampling theorem.

Problem 6.1. Let $a \geqslant 0$ and M be any continuous martingale with $M(0) = x \in (0, a)$. Let τ be the first time M hits either 0 or a. Compute $\boldsymbol{P}(M(\tau) = a)$ and your expected return $\boldsymbol{E}M(\tau)$.