On Monday, we define the idea of set and the operation over it. Intuitively, we think of the element of sets as an **outcome** and think of a set as an **event**. With these notion, further we want to design **chance mechanism** on them to describe real world.

Here comes the important issues:

- What kinds of domain (event space) for us to work on?

- What kind of chance mechanism (probability function) can we choose?

The critical concepts here is that we want to define **axioms** for these two mathematical objects. Moreover, we want such axioms to be

- Complete: Close under any possible operations.

- Compact: Less number of axioms as possible.

1 Recall

1.1 limsup and liminf

Intuitively, we think of sup and inf as:

- sup: Upper bound. In set theory, we think of it as **union** since it contains every future event.

- inf: Lower bound. In set theory, we think of it as **intersection** since every future event contains it.

As a result, limsup and liminf is analogous to the **asymptotic** upper bound and lower bound for a sequence of sets.

1.2 Distributive law & Partition

1. Consider events A, B and sample space Ω,

$$B \cap \Omega = B \cap (A \cup A^{C}) = (B \cap A) \cup (B \cap A^{C})$$

As the sets in the union clause are a partition, distributive law tells us that we can divide the event B into two parts. Here we consider a finite partition so you might not get the feeling. But imagine if the partition is infinite...?
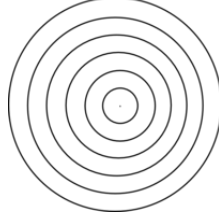
2. Consider,

$$A \cup B = (A \cup B) \cap \Omega = (A \cup B) \cap (A \cup A^C) = A \cup (B \cap A^C) = A \cup (B \backslash A)$$

Geometrically, it's like we union another set $B' = (B \backslash A)$ with A where A and B' are **disjoint**. In a special case: $A_1 \supseteq A_2 \supseteq ...$, it can be think of as a concentric circle And thus we can transform an union into a **disjoint union**.

1.3   General De Morgan's Law

The De Morgan's law holds true as the number of sets is infinite! Formally, let $\{A_\alpha : \alpha \in \Gamma\}$, where Γ is an index set and can be infinite, e.g. real line, or time segment. With such notation, we can define something like

$$\bigcup_{\alpha \leq t} A_\alpha$$

Which can be the events that happen before time t. Then, as De Morgan's law holds true, we can flip the union of events and consider the events that haven't show up.

Intuitively we can think of this as the **history** or **filtration** in a *period sense*. And De Morgan's law provides a convenient way for us to do the computation.

2 σ-Algebra

2.1   Intuitions and basic axioms

By definition, σ-algebra is a collection of subsets of sample space. Intuitively, it is the space of meaningful events. As we mention the **space** of events, σ-algebra must have some nice properties such as closeness under certain operations etc. But here what we concern is not only about what properties (axioms) a σ-algebra has to obey, but also about what's the most compact axioms we can have to capture the sufficient and necessary properties a σ-algebra should follow. And the answer is, there are three of axioms: Consider a σ-algebra $\mathcal{A}$ over sample space Ω,

1. (contain empty event)$\emptyset \in \mathcal{A}$

2. (close under complementation)$\forall A \in \mathcal{A}, A^C \in \mathcal{A}$

3. (close under countable union)$\forall i \in \Gamma \ countable, \ A_i \in \mathcal{A}, \ \bigcup_{i \in \Gamma} A_i \in \mathcal{A}$

As a quick remark, we can see that with these axioms, we also have

- $\Omega \in \mathcal{A}$ (by 1,2)

- Close under countable intersection (by 2, and De Morgan's law)

2.2 Properties

After defining the basic axioms for all σ-algebra, now a more applied issue arises:

$$What's\ the\ smallest\ \sigma - algebra?$$

With some observation, we can easily find out the largest and smallest σ-algebra over sample space Ω is

- Trivial σ-algebra: $\{\emptyset, \Omega\}$

- Total σ-algebra: $\{A : A \subseteq \Omega\}$. That is, the power set of Ω.

To answer this question, we have to clarify some concepts: How do we compare the size of two σ-algebras and to what extent do we want? The first question can be answered with strictly inclusion. If σ-algebra $\mathcal{A}_1$ strictly contains σ-algebra $\mathcal{A}_2$, then we can think of $\mathcal{A}_1$ is larger than $\mathcal{A}_2$.

As to the second question, we can think this in a reverse way: What elements (subsets of Ω) we must have? This question is problem-dependent, so here we imagine that we put all the events that we care into a set A. Then, we got a necessary condition for the σ-algebra that we desire: any σ-algebra that can describe the chance mechanism we're going to use must at least contain all the set in A.

Intuitively, with the above two concepts, it's reasonable to come to a wild guess for the minimal σ-algebra:

$$The\ intersection\ of\ all\ \sigma - algebras\ that\ contains\ A.$$

However, there's still one last thing to check: Is the (countable) intersection of σ-algebras also a σ-algebra? This can be proved in a theorem.

Theorem 1 (σ-algebras are close under countable intersection)
If $\mathcal{A}_\alpha$ is a σ-algebra $\forall \alpha \in \Gamma$, where Γ is a countable index set. Then $\bigcap_{\alpha \in \Gamma} A_\gamma$ is also a σ-algebra.

Proof: We should check three things and it's quite trivial. I believe I will never forget so here I leave it blank. ■

3 Probability Function

Finally, it's time for us to construct the **chance mechanism**! That is, we want to design probability function that maps from a σ-algebra to $[0, 1]$. Moreover, such probability function should be analogous to real life. Namely, not every function from σ-algebra $\mathcal{A}$ to $[0, 1]$ can be out candidate. As a result, we want to design rules (axioms) for probability function in advance. And just as the axioms for σ-algebra, we hope such axioms can be both compact and complete.

Komolgorov Axioms Consider sample space Ω and a σ-algebra $\mathcal{A}$ over it. Then, a probability function P should satisfy the following axioms:

- (Lower bound) $P(A) \geq 0, \forall A \in \mathcal{A}$

- (Upper bound) $P(\Omega) = 1$

- (Countably additive) If $A_\alpha \in \mathcal{A}, \forall \alpha \in \Gamma$, where Γ is a countable **mutually exclusive** index set. Then, $P(\bigcup_{\alpha \in \Gamma} A_\alpha) = \sum_{\alpha \in \Gamma} P(A_\alpha)$

Note that this brings us from set functions to **value function**. That is, we no longer playing only with set, but also take account the number, i.e. the probability.

Formally speaking, we say $(\Omega, \mathcal{A}, P)$ is a probability space if $\mathcal{A}$ is a σ-algebra over Ω and P is a probability function.

It's no hard to see that the most important axioms in Komolgorov axioms is countably additive. It allows us to deal with a countably infinite sequence of events. But you must wonder, can we relax the countably additive to finitely additive? If so, do we need to add other constraints? The answer is positive. In fact, we have

$$\text{Countably additive} = \text{Finitely additive} + \text{Axiom of continuity}$$

Here, we state the axiom of continuity without proof (left as exercise)

Axiom of continuity If a sequence of events $A_1 \supseteq A_2 \supseteq ...$ and $\lim_{n \to \infty} A_n = \emptyset$, then $\lim_{n \to \infty} P(A_n) = 0$

After defining the axioms of probability function, now we want to design a probability function for our own usage. However, here comes the problem, how to assign probability function for a given usage? A simple theorem tells us that it's sufficient to define an unique probability function with the known probability assignments on **trivial partition**.

Theorem 2 *Let sample space $\Omega = \{w_1, w_2, ...\}$, assign probability value to each outcome as $p_1, p_2, ... \geq 0, \sum_i p_i = 1$ respectively. Define the probability function P as*

$$P(A) := \sum_{i:w_i \in A} p_i$$

With some verification to the axioms of probability function, we can see that P is an unique probability function over Ω and the σ-algebra generated by partition $\{w_1, w_2, ...\}$.

Intuitively, once we want to design a probability function for our chance mechanism, it's sufficient to assign probability values over a sufficient partition.

4 Some random notes

4.1 Categorical classification/regression

Suppose we are analyzing data with label set $\{A, B, C\}$ where these labels have no numerical meaning. What can we do? There's a simple solution called: **dummy variable**

The idea is quite simple. Originally, we have a variable $s \in \{A, B, C\}$. But we cannot use s in any regression or numerical model. Now, let's change a way of thinking. We no longer focus on **which** label does s take. Instead, we consider

$$\text{Is } s \text{ takes label } A?$$
$$\text{Is } s \text{ takes label } B?$$
$$\text{Is } s \text{ takes label } C?$$

Then, we generate **two** dummy variables to **quantify** the above intuition.

$$x_1 := \begin{cases} 1, s = A \\ 0, else \end{cases}$$
$$x_2 := \begin{cases} 1, s = B \\ 0, else \end{cases}$$

Now, we have two meaning dummy variables that can be fitted into a numerical model!

4.2 Boole's inequality

Boole's inequality, a.k.a union bound, provides a simple but useful tools when dealing with asymptotic issues.

Theorem 3 (Boole's inequality) *Consider set of events $\{A_\alpha : \alpha \in \Gamma\}$. Then we have*

$$P(\bigcup_{\alpha \in \Gamma} A_\alpha) \leq \sum_{\alpha \in \Gamma} P(A_\alpha)$$

From Boole's inequality, we can derive Borel Cantelli lemma. And from Borel Cantelli lemma, we can define the concept of almost surely convergence.