XIAOYU CHEN

NOTES FOR SAMPLING

Contents

1
Topics Related to Markov Chain

1.1 Time Reversible Markov Chain

Recall that we usuall denote a Markov Chain by $\mathcal{M} = (\Omega, P)$.

Definition 1.1.1. *Reversed Markov Chain*

A reversed MC $\mathcal{M}^R : (\Omega, Q)$ is induced by the following rule:

$$Q(i,j) = \Pr[X_m = j | X_{m+1} = i]$$

Then we have:

$$\begin{aligned} Q(i,j) &= \Pr[X_m = j | X_{m+1} = i] \\ &= \frac{\Pr[X_m = j \wedge X_{m+1} = i]}{\Pr[X_{m+1} = i]} \\ &= \frac{\Pr[X_m = j] P(j,i)}{\Pr[X_{m+1} = i]} \\ &= \frac{\pi(j) P(j,i)}{\pi(i)} \end{aligned}$$

Definition 1.1.2. *time-reversible Markov Chain*

A MC $\mathcal{M} : (\Omega, P)$ is time-reversible, if it equals to its reversed MC $\mathcal{M}^R : (\Omega, Q)$, i.e. $P = Q$.

Note that, this definition equals to saying:

$$\pi(i) P(i,j) = \pi(j) P(j,i), \quad \forall i, j \in \Omega$$

Moreover, if

$$\sum_{\omega \in \Omega} \pi(\omega) = 1$$

, then we could infer that π is a stationary distribution of $\mathcal{M}$.

Lemma 1.1.1.

Once we have

$$\pi(x) P(x,y) = \pi(y) P(y,x)$$

we have

$$\pi(x) P^t(x,y) = \pi(y) P^t(y,x)$$

Proof. We prove this by induction, assume that

$$\pi(x) P^t(x,y) = \pi(y) P^t(y,x)$$

Then

$$\begin{aligned}
\pi(x)P^{t+1}(x,y) &= \sum_z \pi(x)P^t(x,z)P(z,y) \\
&= \sum_z \pi(z)P^t(z,x)P(z,y) \\
&= \sum_z \pi(y)P(y,z)P^t(z,x) \\
&= \pi(y)P^{t+1}(y,x) \qquad \square
\end{aligned}$$

1.2 How to Use Markov Chain to Simulate a Specific Distribution

Suppose we have a specific distribution π, and we want to simulate it using a Markov Chain $\mathcal{M}$. It seems like a pretty hard job, but actually we only need to make sure that:

$$P(X_0, X_1) = c \min\{1, \pi(X_1)/\pi(X_0)\}$$

and

$$P(X_1, X_0) = c \min\{1, \pi(X_0)/\pi(X_1)\}$$

Here, c refers to the same constant in these 2 expression.

where c is some constant less than 1. Suppose $\pi(X_0) \le \pi(X_1)$, it easy to find: $P(X_0, X_1) = c$ and $P(X_1, X_0) = c\pi(X_0)/\pi(X_1)$. And thus:

By symmetry, we only need to consider the case $\pi(X_0) \le \pi(X_1)$.

$$\pi(X_0)P(X_0, X_1) = c\pi(X_0) = \pi(X_1)P(X_1, X_0)$$

which infers that π is a stationary distribution of $\mathcal{M}$.

1.3 Some Useful Tricks for Using Transition Matrix

Oberve that, in the next step, we must in some state in the MC.

Proposition 1.3.1. *Normality*

$$\sum_{y\in\Omega} P(x,y) = 1, \qquad \forall x \in \Omega$$

Then, consider the multiplication between the transition matrix and some functino $f : \Omega \to \mathbb{R}$.

Definition 1.3.1. *Left Multiplication*

$$[fP](y) = \sum_{x\in\Omega} f(x)P(x,y)$$

What's the meaning of the left multiplication?

Proposition 1.3.2. *It Generates A New Distribution*

If f is some distribution, then fP forms a new distribution f'.

Moreover, this could be easily generalized to the t-step case. That is, fP^t is a new distribution.

Proof.

$$\begin{aligned}\sum_{y\in\Omega}[fP](y) &= \sum_{y\in\Omega}\sum_{x\in\Omega} f(x)P(x,y) \\ &= \sum_{x\in\Omega} f(x)\sum_{y\in\Omega} P(x,y) \\ &= \sum_{x\in\Omega} f(x) \\ &= 1\end{aligned}$$

□

Actually this new distribution $f'(x)$ means the probability that select a state x_0 in the distribution of f, move according to the markov chain, and finally end in state x.

Definition 1.3.2. *Right Multiplication*

$$[Pf](x) = \sum_{y\in\Omega} P(x,y)f(y)$$

And, what's the meaning of the right multiplication?

Proposition 1.3.3. *It Could Represent The Probability of move from state x to Some Fixed State Set*

Here, f is no longer a distribution. Suppose here is a state set $A \in \Omega$. Then, we could construct f as a indicator of A:

$$f(x) = \begin{cases} 1, & \text{if } x \in A \\ 0, & \text{otherwise} \end{cases}$$

Then we have:

$$[Pf](x) = \sum_{y\in A} P(x,y) = P(x,A)$$

Note that this could be generalized to t-step case easily. That is

$$[P^t f](x) = P^t(x,A)$$

Note that $||P^t(x,\cdot) - \pi||_{TV}$ is closely related to the mixing time of the markov chain.

1.4 Understand Path Coupling

When I am focusing on a exercise of chapter 7 of Jerrum Book, I find that I have nearly forgot all the details about coupling. To avoid this kind of things in the future, I tend to write this note.

First, what is coupling?

Definition 1.4.1. *Coupling*

Suppose we have a MC Z_t with state space Ω. Then, a coupling for Z_t is an MC (X_t, Y_t) on $\Omega \times \Omega$, with transition probabilities defined by:

$$\Pr[X_1 = x' | X_0 = x, Y_0 = y] = P(x, x')$$
$$\Pr[Y_1 = y' | X_0 = x, Y_0 = y] = P(y, y')$$

Which means X_t and Y_t seems independent in their own perspective.

Here is the main intuition that we could use coupling.

Lemma 1.4.1. *Coupling Lemma*

Let (X_t, Y_t) be any coupling based on Z_t, satisfying Definition 1.4. Suppose $t : [0, 1] \to \mathbb{N}$ is a function satisfying the condition: for all $x, y \in \Omega$, and all $\epsilon > 0$:

$$\Pr[X_{t(\epsilon)} \neq Y_{t(\epsilon)} | X_0 = x, Y_0 = y] \leq \epsilon$$

Then the mixing time of Z_t is bounded by $t(\epsilon)$.

Note that this condition should be satisfied for any $x, y \in \Omega$. It is not easy to reach these requirements.

Proof. For any $x \in \Omega$ and any $A \subseteq \Omega$, we have

$$\begin{aligned} P^t(x, A) &= \Pr[x_t \in A] \\ &\geq \Pr[X_t = Y_t \wedge Y_t \in A] \\ &= 1 - \Pr[X_t \neq Y_t \vee Y_t \notin A] \\ &\geq 1 - (\Pr[X_t \neq Y_t] + \Pr[Y_t \notin A]) \\ &\geq \Pr[Y_t \in A] - \epsilon \\ &= \pi(A) - \epsilon \end{aligned}$$

□

Definition 1.4.2. *Adjacent States*

Suppose we have a MC Z_t. Then, two state $x, y \in \Omega$ are adjacent if

$$P(x, y) > 0$$

Which means state x could be translate to state y within one iteration of Zt.

Here, **Path Coupling** allows us to design coupling only between adjacent pairs.

Lemma 1.4.2. *Neighbor to Global*

Suppose we have a coupling (X_t, Y_t) based on Z_t for adjacent pairs. And for each adjacet state X_0 and Y_0 in Z_t,

$$\mathbb{E}[d(X_1, Y_1)|X_0, Y_0] \leq \varrho d(X_0, Y_0) \tag{1.1}$$

Then, the Inequality 1.1 could be extended to all pairs of states.

Proof. Suppose we have two states $x_0, y_0 \in \Omega$ arbitary. We extend our **local coupling** (X_t, Y_t) to a **global coupling** like this:

```
$GC(x_0, y_0)$ begin
    /* find a shortest path from $x_0$ to $y_0$ where */
    /* $x_0 = z^{(0)}, z^{(1)}, \cdots, z^{(l)} = y_0$ */
    $Z^{(0)} \leftarrow Z_t\{z^{(0)}\}$ // $Z_t$ is the original MC
    for $i = 1 \to l$ do
        do
            $(Z^*, Z^{(i)}) \leftarrow (X_t, Y_t)\{z^{(i-1)}, z^{(i)}\}$
        until $Z^* = Z^{(i-1)}$;
        /* restrict the first element to be $Z^{(i-1)}$ */
    end
    return $(x_1 \leftarrow Z^{(0)}, y_1 \leftarrow Z^{(l)})$
end
```

Then we only need to show that our **global coupling** *GC* works as we expected. First note that $Z^{(0)}$ is select from the distribution $P(x_0, \cdot)$. And moreover

$$\begin{aligned}
P(Z^{(1)}) &= \sum_{Z^{(0)} \in \Omega} \Pr(z^{(0)}, Z^{(0)}) \Pr[(z^{(0)}, z^{(1)}) \mapsto (Z^{(0)}, Z^{(1)}) | z^{(0)} \mapsto Z^{(0)}] \\
&= \sum_{Z^{(0)} \in \Omega} P(z^{(0)}, Z^{(0)}) \frac{\Pr[(z^{(0)}, z^{(1)}) \mapsto (Z^{(0)}, Z^{(1)})]}{\sum_{Z \in \Omega} \Pr[(z^{(0)}, z^{(1)}) \mapsto (Z^{(0)}, Z^{(1)})]} \\
&= \sum_{Z^{(0)} \in \Omega} P(z^{(0)}, Z^{(0)}) \frac{\Pr[(z^{(0)}, z^{(1)}) \mapsto (Z^{(0)}, Z^{(1)})]}{P(z^{(0)}, Z^{(0)})}, \quad \text{from the definition of a coupling} \\
&= \sum_{Z^{(0)} \in \Omega} \Pr[(z^{(0)}, z^{(1)}) \mapsto (Z^{(0)}, Z^{(1)})] \\
&= P(z^{(1)}, Z^{(1)}), \quad \text{from the definition of a coupling}
\end{aligned}$$

Then, by induction, we know that $Z^{(i)}$ was chose from the distribution

$$P(z^{(i)}, \cdot)$$

So we know that *GC* is a coupling. And by the linearity of expectation, we have

$$\begin{aligned}
\mathbb{E}[d(x_1, y_1) | x_0, y_0] &\le \sum_{i=0}^{l-1} \mathbb{E} d(Z^{(i)}, Z^{(i+1)}) \\
&\le \varrho \sum_{i=0}^{l-1} d(z^{(i)}, z^{(i+1)}) \\
&= \varrho d(x_0, y_0), \quad \text{since we choose a shortest path}
\end{aligned}$$

The reason why that we could calcuate each pairs expected distance by the expression above is not so trivial. It turns out that we could treat each edge $(Z^{(i)}, Z^{(i+1)})$ as it is chose by our **local coupling**. And it has no side effects with other edges. So its expected length could bound by Inequality 1.1. □

1.5 Understand Canonical Path And Flow

After reading the topic on canoical path on Jerrum Book, I do not think that I have a very deeply understanding of this technique. Because the calculation used by it is very complex and magical, where the book does not explain the intuition behind the calculation

Definition 1.5.1. *Indicator Function*

The function f we use here, should be an indicator of some set $A \subseteq \Omega$. That is

$$f(x) := [x \in A], \quad \forall x \in \Omega$$

In the latter part of this article, we may use f_A instead of f if we want to refer to specific set A.

Definition 1.5.2. *Inner Product On Distribution*

Suppose we have two function (or vector) f and g, the we could define their inner product based on π:

$$\langle f, g \rangle_\pi := \sum_{x \in \Omega} f(x) g(x) \pi(x)$$

Lets have a look at the following concept with our new tools:

Fact 1.5.1.

$$\mathbb{E}_\pi f = \pi(A)$$

Proof. Lets treat f and π as two vector, then:

$$\begin{aligned} \mathbb{E}_\pi f &= \pi f \\ &= \pi([x \in A]) \\ &= \pi(A) \end{aligned}$$

□

Corollary 1.5.1.

Suppose P is a transition matrix of some Markov Chain, then:

$$\mathbb{E}_\pi[Pf] = \mathbb{E}_\pi f = \pi(A)$$

Proof.

$$\begin{aligned}\mathbb{E}_\pi[Pf] &= \pi P f \\ &= \pi f, \quad \text{since } \pi \text{ is stationary distribtion} \\ &= \mathbb{E}_\pi\end{aligned}$$ □

A interesting thing here is that we could treat a Markov Chain with a stationary distribution π as a multicommodity flow network, although the "flow" we use here is different from the flow in context of algorithm. This network works under the following rules:

```
Network Working Schedule
begin
    $t = 0$ // Time Counter
    $P$ // The Transition Matrix of Our MC
    $\pi_0$ // Some Distribution
    /* $\pi_0(v)$ means the quantity of commodities on $v$    */
    while $t \leftarrow t + 1$ do
        $\pi_t(\cdot) \leftarrow 0$
        for $u, v \in \Omega$ do
            $\pi_t(v) \leftarrow \pi_t(v) + \pi_{t-1}(u)P(u,v)$
            /* $\pi_t(v) = \sum_{u \in \Omega} \pi_{t-1}(u)P(u,v)m,$    */
        end
    end
end
```

It is clear that from time $t-1$ to time t, there is a $\pi_{t-1}(u)P(u,v)$ units of flow on edge (u,v). And it is easy to verify that from time $t-1$ to time t, there are

$$\sum_{u \in \Omega} \pi_{t-1}(v)P(v,u) = \pi_{t-1}$$

units of commodities flow out v and there are

$$\sum_{u \in \Omega} \pi_{t-1}(u)P(u,v) = \pi_t$$

units of commodities flow in v.

So, it seems that any vertex in the network will run out of all its commodities in current turn, and get all its commodities for the next turn from its neighbors. To go further in this topic, we borrow some notations from the textbook:

Definition 1.5.3. *Flow Function In One Step*

We want a functoin $Q_1^\pi(\cdot,\cdot)$ to measure the quantity of the flow from one part to another part in one turn (start with the distribution π). Suppose we have two sets A and B, then the flow from A to B is represent by:

$$Q_1^\pi(A,B) = \sum_{a\in A, b\in B} \pi(a)P(a,b)$$

Here, π could be any distribution. If we need to distinguish the transition matrix P in the context, we use ${}_PQ_1^\pi$ instead of Q_1^π.

Now consider flow cross many steps.

Proposition 1.5.1.

Suppose the commodities at each vertex will be well mixed after each turn, then

$$\mathbb{E}[Q_t^\pi(A,B)] = \sum_{a\in A, b\in B} \pi(a)P^t(a,b)$$

Here, π could be any distribution.

Proof. We prove this by induction.

In the base case: $\mathbb{E}[Q_1^\pi(A,B)] = Q_1^\pi(A,B)$.

In the inductive case: Suppose $t = n$, and our assumption holds in $t < n$. Then

$$\begin{aligned}
\mathbb{E}[Q_n^\pi(A,B)] &= \sum_{a\in A, b\in B} \mathbb{E}[Q_n^\pi(a,b)] \\
&= \sum_{\substack{a\in A, b\in B\\ m\in\Omega}} \mathbb{E}[Q_{n-1}^\pi(a,m)]P(m,b) \qquad (1.2) \\
&= \sum_{\substack{a\in A, b\in B\\ m\in\Omega}} \pi(a)P^{n-1}(a,m)P(m,b), \quad \text{by assumption} \\
&= \sum_{a\in A, b\in B} \pi(a) \sum_{m\in\Omega} P^{n-1}(a,m)P(m,b) \\
&= \sum_{a\in A, b\in B} \pi(a)P^n(a,b)
\end{aligned}$$

Because we could treat $P(a,\cdot)$ as a distribution. So the expect quantity of flow (which has moved from a to m) moves from m to b in step n is

$$\mathbb{E}[Q_{n-1}^\pi(a,m)]P(m,b)$$

Actually, we could treat the commodities as some liquid distinguishing by their start point. In each turn, we mix them together first and then push them to pipes (or edges) by the ratio ($P(u,v)$) on the pipe. (Maybe we could throw the $\mathbb{E}$ away by understanding the process like this)

And hence we have the Equation 1.2. □

Corollary 1.5.2.

$$\lim_{t\to\infty} \mathbb{E}[Q_t^\mu(A,B)] = \mu(A)\pi(B)$$

Proof.

$$\begin{aligned}\lim_{t\to\infty} \mathbb{E}[Q_t^\mu(A,B)] &= \lim_{t\to\infty} \sum_{a\in A, b\in B} \mu(a)P^t(a,b)\\ &= \sum_{a\in A, b\in B} \mu(a)\pi(b),\ \pi \text{ is the stationary distribution}\\ &= \mu(A)\pi(B)\end{aligned}$$

□

For convenience, we use Q_t instead of $\mathbb{E}_\pi[Q_t]$ latter in this article.

Definition 1.5.4.

For convenience:

$$Q^\mu(A,B) := \lim_{t\to\infty} \mathbb{E}[Q_t^\mu(A,B)] = \mu(A)\pi(B)$$

Moreover,

$$Q(A,B) := \lim_{t\to\infty} \mathbb{E}[Q_t^\pi(A,B)] = \pi(A)\pi(B)$$

where π is the stationary distribution of the MC.

So far so good, we have a useful tool to explain the intuition behind the calculation for canonical path on Jerrum Book. Now, we need to focus on some important concepts which is used in the proof of the "canonical path" (e.g. $\text{Var}_\pi f$, $\text{Var}_\pi[P_{zz}f]$, $\mathcal{E}_P(f,f)$, etc.).

1.5.1 *Explaination For* $\text{Var}_\pi f$

First, we should note that there is a special form of variance:

$$\text{Var}_\pi f = \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)(f(x)-f(y))^2$$

Note that the distribution π here is the stationary distribution for the MC.

where you could refer to Proposition 2.2.1 for how this comes.

Lemma 1.5.1.

$$\text{Var}_\pi f = \pi(A^c)\pi(A), \quad \textit{for } f = f_A$$

Proof. Note that

$$(f(x)-f(y))^2 = [x \in A] \oplus [y \in A]$$

So we have

$$\begin{aligned}
\text{Var}_\pi f &= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)(f(x)-f(y))^2 \\
&= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)([x\in A]\oplus[y\in A]) \\
&= \frac{1}{2}\sum_{x\in A, y\in\Omega\setminus A} \pi(x)\pi(y) + \frac{1}{2}\sum_{x\in\Omega\setminus A, y\in A} \pi(x)\pi(y) \\
&= \sum_{x\in A, y\in\Omega\setminus A} \pi(x)\pi(y) \\
&= \sum_{x\in A, y\in A^c} \pi(x)\pi(y) \\
&= \pi(A^c)\pi(A) \qquad \square
\end{aligned}$$

Corollary 1.5.3.

$$\text{Var}\pi f + (\mathbb{E}_\pi f)^2 = \pi(A)$$

Proof. Note that

$$(\mathbb{E}_\pi f)^2 = \pi(A)\pi(A)$$

So we have

$$\begin{aligned}
\text{Var}\pi f + (\mathbb{E}_\pi f)^2 &= \pi(A^c)\pi(A) + \pi(A)\pi(A) \\
&= \pi(\Omega)\pi(A) \\
&= \pi(A) \qquad \square
\end{aligned}$$

Lemma 1.5.2.

$$\text{Var}_\pi f + (\mathbb{E}_\pi f)^2 = \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)P(x,y)(f(x)^2 + f(y)^2)$$

This equation appears on the book, we could explain it use the language of flow.

Proof. According to the rule of the flow. $\pi(A)$ is the quantity of the commodities in the begining. Because π is the stationary disbtribution of MC, so $\pi P = \pi$. This means the quantity of commodities in A

should still be $\pi(A)$ after one trun of the flow. So we have

$$\begin{aligned}
\pi(A) &= \text{\# of commodities from } A + \text{\# of commodities from } A^c \\
&= \sum_{u\in A, v\in A} \pi(u)P(u,v) + \sum_{u\in A^c, v\in A} \pi(u)P(u,v) \\
&= \sum_{[u\in A \wedge v\in A]} \pi(u)P(u,v) + \sum_{[u\in A^c \wedge v\in A]} \pi(u)P(u,v) \\
&= \sum_{[u\in A \wedge v\in A]} \pi(u)P(u,v) + \frac{1}{2} \sum_{[u\in A]\oplus[v\in A]} \pi(u)P(u,v) \\
&= \frac{1}{2} \sum_{x,y\in\Omega} \pi(x)P(x,y)(f(x)^2 + f(y)^2)
\end{aligned}$$

The last step is because $\frac{1}{2}(f(x)^2 + f(y)^2)$ could be treat as an indicator:

$$\frac{1}{2}(f(x)^2 + f(y)^2) = \begin{cases} 1, & x \in A \wedge y \in A \\ 1/2, & [x \in A] \oplus [y \in A] \\ 0, & \text{otherwise} \end{cases} \qquad \square$$

1.5.2 *Explaination For $\mathcal{E}_P(f,f)$*

Fact 1.5.2.

$$\mathcal{E}_P(f,f) = {}_PQ_1^\pi(A, A^c)$$

So $\mathcal{E}_P(f,f)$ means the flow from A^c to A in one step (start from π).

Proof. From Jerrum Book, we could know:

$$\begin{aligned}
\mathcal{E}_\pi(f,f) &= \frac{1}{2} \sum_{x,y\in\Omega} \pi(x)P(x,y)(f(x) - f(y))^2 \\
&= \frac{1}{2} \sum_{x,y\in\Omega} \pi(x)P(x,y)[x \in A] \oplus [y \in A] \\
&= \frac{1}{2} \sum_{[x\in A]\oplus[y\in A]} \pi(x)P(x,y) \\
&= \sum_{x\in A^c, y\in A} \pi(x)P(x,y) \\
&= Q_1^\pi(A^c, A)
\end{aligned} \qquad \square$$

Fact 1.5.3.

$$\mathrm{Var}_\pi f + (\mathbb{E}_\pi f)^2 - \frac{1}{2}\mathcal{E}_P(f,f) = \frac{1}{4} \sum_{x,y\in\Omega} \pi(x)P(x,y)(f(x) + f(y))^2$$

Proof. First we have:

$$\frac{1}{4}(f(x)+f(y))^2 = \begin{cases} 1, & x \in A, y \in A \\ 1/4, & x \in A, y \in A^c \\ 1/4, & x \in A^c, y \in A \\ 0, & \text{otherwise} \end{cases}$$

So

$$\begin{aligned} &\mathrm{Var}_\pi f + (\mathbb{E}_\pi f)^2 - \frac{1}{2}\mathcal{E}_P(f,f) \\ &= \sum_{x,y \in A} \pi(x)P(x,y) + \frac{1}{2}\sum_{x \in A^c, y \in A} \pi(x)P(x,y) \\ &= Q_1^\pi(A,A) + \frac{1}{2}Q_1^\pi(A^c,A) \\ &= \pi(A) - \frac{1}{2}Q_1^\pi(A^c,A) \end{aligned}$$

□

Corollary 1.5.4.

$$\mathrm{Var}_\pi f + (\mathbb{E}_\pi f)^2 - \mathcal{E}_{P_{zz}}(f,f) = \pi(A) - {}_{P_{zz}}Q_1^\pi(A^c,A) = {}_{P_{zz}}Q_1^\pi(A,A)$$

1.5.3 *Explaination For* $\mathrm{Var}_\pi[P_{zz}f]$

Since $\mathbb{E}_\pi[P_{zz}f] = \mathbb{E}_\pi f$, we only consider $\mathrm{Var}_\pi[P_{zz}f] + (\mathbb{E}_\pi f)^2$.

Lemma 1.5.3.

$$\mathrm{Var}_\pi[P_{zz}f] + (\mathbb{E}f)^2 = {}_{P_{zz}}Q_2^\pi(A,A)$$

Proof.

$$\begin{aligned} \mathrm{Var}_\pi[P_{zz}f] + (\mathbb{E}f)^2 &= \sum_{x \in \Omega} \pi(x)([P_{zz}f](x))^2 \\ &= \sum_{x \in \Omega} \pi(x)(P_{zz}(x,A))^2 \\ &= \sum_{x \in \Omega} \pi(x) \sum_{y \in A} P_{zz}(x,y) \sum_{z \in A} P_{zz}(x,z) \\ &= \sum_{\substack{y \in A, z \in A \\ x \in \Omega}} \pi(x)P_{zz}(x,y)P_{zz}(x,z) \\ &= \sum_{\substack{y \in A, z \in A \\ x \in \Omega}} \pi(y)P_{zz}(y,x)P_{zz}(x,z), \quad \text{assume } \pi(x)P_{zz}(x,y) = \pi(y)P_{zz}(y,x) \\ &= \sum_{y \in A, z \in A} \pi(y)P_{zz}^2(y,z) \\ &= {}_{P_{zz}}Q_2^\pi(A,A) \end{aligned}$$

□

Corollary 1.5.5.

$$\mathrm{Var}_\pi[P_{zz}^t f] + (\mathbb{E}_\pi P_{zz}^t f)^2 = {}_{P_{zz}}Q_{2t}^\pi$$

We find the connection between $\mathrm{Var}_\pi f$ and $\mathrm{Var}_\pi[P_{zz} f]$, since Corollary 1.5.4

Theorem 1.5.1.

$$\mathrm{Var}_\pi(P_{zz} f) \le \mathrm{Var}_\pi f - \mathcal{E}_{P_{zz}}(f, f)$$

This theorem is true on any $f : \Omega \to \mathbb{R}$ (i.e. not just indicator).

If you want to relax the restriction on f, you need to proof this theorem in the way the Jerrum Book does. But that is pretty hard to understand. The proof we give here only aims at gaining the intuition.

Proof. Since

$$\begin{aligned} \mathrm{Var}_\pi(P_{zz} f) + (\mathbb{E}_\pi f)^2 &\le \mathrm{Var}_\pi f + (\mathbb{E}_\pi f)^2 - \mathcal{E}_{P_{zz}}(f, f) \\ {}_{P_{zz}}Q_2^\pi(A, A) &\le {}_{P_{zz}}Q_1^\pi(A, A) \end{aligned} \tag{1.3}$$

Thus we only need to proof Equation 1.3. Actually, under the assumption of $\pi(x)P(x, y) = \pi(y)P(y, x)$, we have:

$$\begin{aligned} {}_{P_{zz}}Q_2^\pi(A, A) &= \sum_{x \in \Omega} \pi(x) P_{zz}(x, A) P_{zz}(x, A) \\ &\le \sum_{x \in \Omega} \pi(x) P_{zz}(x, A), \quad \text{Since } P_{zz}(x, A) \le 1 \\ &= {}_{P_{zz}}Q_1^\pi(A, A) \end{aligned}$$

□

Corollary 1.5.6.

$$\pi(A)\pi(A) \le {}_{P_{zz}}Q_{t+1}^\pi(A, A) \le {}_{P_{zz}}Q_t^\pi(A, A) \le {}_{P_{zz}}Q_1^\pi(A, A)$$

Theorem 1.5.2.

For $P^t = P_{zz}^t$, we have

$$\mathrm{Var}_\pi[P^{t+1} f] \le \mathrm{Var}_\pi[P^t f] - \mathcal{E}_{P_{zz}}(P^t f, P^t f)$$

Proof. First, note that

$$\mathrm{Var}_\pi[P^{t+1} f] + (\mathbb{E}_\pi[P^{t+1} f])^2 = Q_{2t+2}^\pi(A, A)$$

and

$$\mathrm{Var}_\pi[P^t f] + (\mathbb{E}_\pi[P^t f])^2 = Q_{2t}^\pi(A, A)$$

Thus, we only need to prove that

$$\mathcal{E}_{P_{zz}}(P^t f, P^t f) \leq Q^\pi_{2t} - Q^\pi_{2t+2}$$

Moreover

$$\begin{aligned}
\mathcal{E}_{P_{zz}}(P^t f, P^t f) &= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t f(x) - P^t f(y))^2 \\
&= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A) - P^t(y,A))^2 \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - \sum_{x,y\in\Omega} \pi(x)P(x,y)P^t(x,A)P^t(y,A) \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - \sum_{x,y\in\Omega} \pi(x)P(x,y)\sum_{a\in A} P^t(x,a)\sum_{b\in A} P^t(y,A) \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - \sum_{x,y\in\Omega} P(x,y)\sum_{a\in A} \pi(a)P^t(a,x)\sum_{b\in A} P^t(y,A) \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - \sum_{\substack{x,y\in\Omega \\ a,b\in A}} \pi(a)P^t(a,x)P(x,y)P^t(y,b) \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - \sum_{a,b\in A} \pi(a)P^{2t+1}(a,b) \\
&= \sum_{x,y\in\Omega} \pi(x)P(x,y)(P^t(x,A))^2 - Q^\pi_{2t+1}(A,A) \\
&\leq \sum_{x,y\in\Omega} \pi(x)(P^t(x,A))^2 - Q^\pi_{2t+1}(A,A) \\
&= Q^\pi_{2t}(A,A) - Q^\pi_{2t+1}(A,A)
\end{aligned}$$

So, we only need to prove that

$$Q^\pi_{2t}(A,A) - Q^\pi_{2t+1}(A,A) \leq Q^\pi_{2t}(A,A) - Q^\pi_{2t+2}(A,A)$$

Since $Q^\pi_{2t+2}(A,A) \leq Q^\pi_{2t+1}(A,A)$, this inequality is obviously true.

□

We won't go futher here because we have already gain some intuition from above. Next, we only need to use the lower bound in the book:

$$\mathcal{E}_P(f,f) \geq \frac{1}{\varrho}\mathrm{Var}_\pi f$$

And we have

$$\mathrm{Var}_\pi(P_{zz}f) \leq (1 - \frac{1}{2\varrho})\mathrm{Var}_\pi f$$

Its enough, and we stop here.

1.6 *Lower Bound of Markov Chains Mixing Time*

This section summaries some techniques to analysis the lower bound of a Markov Chains' mixing time.

1.6.1 Counting Bound

Counting bound is form by a simple intuition:

If the possible locations of a chain after t steps do not form a significant ***fraction*** *of the state space, then the distribution of the chain at time t cannot be close to uniform.*

If we seen the state space $\mathcal{X}$ as a graph G. Then we could give a lower bound for this ***fraction*** *by using the relationship between the size n of the graph and the maximum degree Δ of this graph.*

Let

$$\deg(v) := |\{u|P(v,u) > 0\}|$$

then

$$\Delta = \max_{x\in\mathcal{X}} \deg(x)$$

Definition 1.6.1.

Let $\mathcal{X}_t^x$ be the states that could be reached from x in exactly t steps. Its quite easy for us to note that $|\mathcal{X}_t^x| \leq \Delta^t$.

If $\Delta^t < (1-\varepsilon)|\mathcal{X}|$, and π is the uniform distribution, then we have

$$\begin{aligned} ||P^t(x,\cdot) - \pi||_{TV} &\geq P^t(x,\mathcal{X}_t^x) - \pi(\mathcal{X}_t^x) \\ &= 1 - \frac{\Delta^t}{|\mathcal{X}|} \\ &> \varepsilon \end{aligned}$$

So, we know that if $\Delta^t < (1-\varepsilon)|\mathcal{X}|$, then the total variansion distance does not reach the target ε as we want. Which means we have

$$\begin{aligned} \Delta^{t_{mix}(\varepsilon)} &\geq (1-\varepsilon)|\mathcal{X}| \\ t_{mix}(\varepsilon) &\geq \frac{\log(|\mathcal{X}|(1-\varepsilon))}{\log\Delta} \end{aligned}$$

1.6.2 Diameter Bound

For convenience, we define some notations here.

Definition 1.6.2.

$$\begin{aligned} d(t) &:= \max_{x\in\mathcal{X}} ||P^t(x,\cdot) - \pi(x)||_{TV} \\ \overline{d}(t) &:= \max_{x,y\in\mathcal{X}} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \end{aligned}$$

Lemma 1.6.1.

$$d(t) \leq \overline{d}(t) \leq 2d(t)$$

Proof. *(1) ($\overline{d}(t) \leq 2d(t)$):*

By triangle inequality (Proposition 2.4.2), we have

$$\begin{aligned}\bar{d}(t) &= \max_{x,y\in\mathcal{X}} ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \\ &= ||P^t(x_0,\cdot) - P^t(y_0,\cdot)||_{TV} \\ &\leq ||P^t(x_0,\cdot) - \pi||_{TV} + ||P^t(y_0,\cdot) - \pi||_{TV}, \quad \text{triangle inequality} \\ &\leq \max_{x\in\mathcal{X}} ||P^t(x,\cdot) - \pi||_{TV} + \max_{y\in\mathcal{X}} ||P^t(y,\cdot) - \pi||_{TV} \\ &= 2d(t)\end{aligned}$$

(2) ($d(t) \leq \bar{d}(t)$):

$$\begin{aligned}d(t) &= ||P^t(x,\cdot) - \pi||_{TV} \\ &= \max_{A\subseteq\mathcal{X}} |P^t(x,A) - \pi(A)| \\ &= |P^t(x,A) - \pi(A)|, \quad \text{for some } A \\ &\text{since } \pi P^t = \pi, \text{ we have } \pi(A) = \textstyle\sum_{y\in\mathcal{X}} \pi(y)P^t(y,A) \\ &= \sum_{y\in\mathcal{X}} \pi(y)|P^t(x,A) - P^t(y,A)| \\ &\leq \sum_{y\in\mathcal{X}} \pi(y)||P^t(x,\cdot) - P^t(y,\cdot)||_{TV} \\ &\leq ||P^t(x,\cdot) - P^t(y,\cdot)||_{TV}\end{aligned}$$

□

In a graph G formed by a state space Ω, if $\text{dis}(x,y) = L$, then $P^{\lfloor(L-1)/2\rfloor}(x,\cdot)$ and $P^{\lfloor(L-1)/2\rfloor}(y,\cdot)$ are possible on disjoint vertex sets. More precisely,

$$||P^{\lfloor(L-1)/2\rfloor}(x,\cdot) - P^{\lfloor(L-1)/2\rfloor}(y,\cdot)||_{TV} = 1$$

So if L is the diameter of the whole graph, we have

$$d(\lfloor(L-1)/2\rfloor) \leq \bar{d}(\lfloor(L-1)/2\rfloor) = 1 \leq 2d(\lfloor(L-1)/2\rfloor)$$

So we have

$$d(\lfloor(L-1)/2\rfloor) \geq \frac{1}{2}$$

So, if $\varepsilon < \frac{1}{2}$, by

$$t_{mix}(\varepsilon) = \min\{t : d(t) < \varepsilon\}$$

we have

$$t_{mix}(\varepsilon) \geq \frac{L}{2}$$

1.6.3 *Bottleneck Ratio*

A bottleneck makes portions of $\mathcal{X}$ difficult to reach from some starting locations, limiting the speed of convergence.

Definition 1.6.3. *Bottleneck Ratio*

$$\begin{aligned}
Q(x,y) &:= \pi(x)P(x,y) && \text{edegs measure}\\
Q(A,B) &:= \sum_{x\in A, y\in B} \pi(x)P(x,y) \\
\Phi(S) &:= \frac{Q(S,S^c)}{\pi(S)} && \text{bottleneck ratio of } S\\
\Phi_* &:= \min_{S:\pi(S)\leq\frac{1}{2}} \Phi(S) && \text{bottleneck ratio of whole chain}
\end{aligned}$$

Theorem 1.6.1.

$$t_{mix}(1/4) \geq \frac{1}{4\Phi_*}$$

Proof.

$$\begin{aligned}
P_\pi\{X_0 \in A, X_t \in A^c\} &\leq \sum_{r=1}^{t} P_\pi\{X_{r-1} \in A, X_r \in A^c\}, \quad \text{all } X_r \text{ has distribution } \pi\\
&= tP_\pi\{X_0 \in A, X_1 \in A^c\}\\
&= tQ(A, A^c)\\
P_\pi\{X_t \in A^c | X_0 \in A\} &\leq t\Phi(A)
\end{aligned}$$

So, we have

$$P_\pi\{X_t \in A | X_0 \in A\} \geq 1 - t\Phi(A)$$

So, there exists x whit $P^t(x, A) \geq 1 - t\Phi(A)$. Therefore

$$d(t) \geq 1 - t\Phi(A) - \pi(A)$$

If $\pi(A) \leq 1/2$ and $t < 1/[4\Phi(A)]$, then $d(t) > 1/4$. So

$$t_{mix} < 1/[4\Phi(A)]$$

. Maximizing over A with $\pi(A) \leq 1/2$ completes the proof. □

1.6.4 *Distinguishing Statistics*

Fact 1.6.1.

If X is a random variable on $\mathcal{X}$ with distribution μ, then for some function $f : \mathcal{X} \to \Lambda$, $f(X)$ is a random variable on Λ with distribution μf^{-1}.

Lemma 1.6.2.

Let μ and ν be probability distributions on $\mathcal{X}$, and let $f : \mathcal{X} \to \Lambda$ be a function on $\mathcal{X}$, where Λ is a finite set. Then

$$||\mu - \nu||_{TV} \geq ||\mu f^{-1} - \nu f^{-1}||_{TV}$$

Proof. Obviously, we have

$$\max_{B \subset \Lambda} |\mu f^{-1}(B) - \nu f^{-1}(B)| \leq \max_{A \subset \mathcal{X}} |\mu(A) - \nu(A)|$$

since if $B \subset \Lambda$, then $f^{-1}(B) \subset \mathcal{X}$. □

Theorem 1.6.2.

For $f : \mathcal{X} \to \mathbb{R}$, define $\sigma_*^2 = \max\{\mathrm{Var}_\mu(f), \mathrm{Var}_\nu(f)\}$. If

$$|\mathbb{E}_\mu(f) - \mathbb{E}_\nu(f)| \geq r\sigma_*^2$$

then

$$||\mu - \nu||_{TV} \geq 1 - \frac{8}{r^2}$$

In particular, if for a Markov chain (X_t) with transition matrix P the function f satisfies

$$|\mathbb{E}_x[f(X_t)] - \mathbb{E}_\pi(f)| \geq r\sigma_*^2$$

then

$$||P^t(x, \cdot) - \pi||_{TV} \geq 1 - \frac{8}{r^2}$$

Proof. Assume that $\mathbb{E}_\mu(f) \leq \mathbb{E}_\nu(f)$. Let $A = (\mathbb{E}_\mu(f) + r\sigma_*^2/2, +\infty)$ be an interval. Then by Chebyshev inequality (see Theorem 2.5.1) we have

$$\mu f^{-1}(A) \leq \tfrac{4}{r^2} \text{ and } \nu f^{-1}(A) \geq 1 - \tfrac{4}{r^2}$$

hence

$$||\mu f^{-1} - \nu f^{-1}||_{TV} \geq 1 - \frac{8}{r^2}$$

. And by using the lemma we have proved above, we could finish our proof. □

1.7 *Continuous Time Markov Chain*

Exercise 1.7.1 (from Exercise 20.3 of Markov Chains and Mixing Times)**.** *Let $T_1, T_2, \cdots$ be an i.i.d sequence of exponential random variables of rate μ, let $S_k = \sum_{i=1}^k T_i$, and let $N_t = \max\{k : S_k \leq t\}$.*

(a) Show that S_k has a gamma distribution with shape parameter k and rate parameter μ, i.e. its density function is

$$f_k(s) = \frac{\mu^k s^{k-1} e^{-\mu s}}{(k-1)!}$$

(b) Show by computing $\Pr\{S_k \leq t < S_{k+1}\}$ that N_t is a Possion random variable

Proof. (a): First we know that $f_{T_1}(x) = f_{T_2}(x) = \mu e^{-\mu x}$ when $x \geq 0$. Then

$$\begin{aligned} f_{T_1+T_2}(s) &= \int_{-\infty}^{\infty} f_{T_1}(t) f_{T_2}(s-t)\mathrm{d}t \\ &= \int_0^s \mu e^{-\mu t} \mu e^{-\mu(s-t)}\mathrm{d}t \\ &= \int_0^s \mu^2 e^{-\mu s}\mathrm{d}t \\ &= \mu^2 e^{-\mu s} t\Big|_0^s \\ &= \mu^2 e^{-\mu s} s \end{aligned}$$

And its very easy to generalize this result to the sum of k random variables.

(b): Since S_k and T_{k+1} are two independent random variables, we have:

$$\begin{aligned} \Pr\{S_k \leq t < S_{k+1} &= \Pr\{S_k \leq t < S_k + T_{k+1}\} \\ &= \int_0^t f_k(s)\mathrm{d}s \cdot \Pr\{T_{k+1} > t-s\} \\ &= \int_0^t \frac{\mu^k s^{k-1} e^{-\mu s}}{(k-1)!}\mathrm{d}s \cdot e^{-\mu(t-s)} \\ &= \int_0^t \frac{\mu^k s^{k-1} e^{-\mu t}}{(k-1)!}\mathrm{d}s \\ &= \frac{\mu^k s^k e^{-\mu t}}{k!}\Bigg|_0^t \\ &= \frac{(\mu t)^k e^{-\mu t}}{k!} \end{aligned}$$

□

1.8 Coupling From the Past

Coupling from the past (CFTP) is a method with which we could perform a perfect sampling according to some distribution. I can not find any convincing proof of this result, so I decide to prove it myself.

First, lets view Markov Chain in a new but nature way. Suppose we have a Markov Chain $\mathcal{M}$ defined on some state set $S =$

$\{s_1, s_2, \cdots, s_n\}$. Note that when we execute $\mathcal{M}$, then in each turn, this chain will fix some parameters according to some distribution and move from some states to other states. So, if we fix this random parameters, then $\mathcal{M}$ becomes a determined process, and we could view it as a function $f : S \to S$. For convenience, lets denote the determined process from time i to time $i+1$ by f_i. The pseudocode of CFTP is shown below.

```
Coupling From The Past () begin
    $T \leftarrow 1$
    repeat
        $f \leftarrow f_{-T} \circ f_{-(T-1)} \circ \cdots \circ f_{-1}$
        $T \leftarrow T + 1$
    until $f$ becomes a constant function;
    return $f(s_1)$
end
```

The $\circ$ here means compose operator where $f \circ g(x) = g(f(x))$, i.e. give f's output as g's input.

Here, f_{-i} is fixed (by random) at the first time we encountered it, and will be **reused** later.

The idea of this sampler is quite simple. Suppose we have a Markov Chain $\mathcal{M}$ which starts at $-\infty$ and stops at 0, then its pretty sure that it will have the distribution π at time 0 (no matter which state it starts from). Then, if we randomly determine some last finite steps of the chain (say T steps), we may have the chance to recover the sample of this MC by this steps.

And, its quite easy to see that when the last T steps forms a constant function we could easily recover the sample from this MC. So we have the following theorem:

Theorem 1.8.1.

If we have a Markov Chain $\mathcal{M}$ such that

$$\lim_{t \to \infty} \Pr[f_0 \circ f_1 \circ \cdots \circ f_t \text{ is a constant function}] = 1$$

where $f_0, f_1, \cdots, f_t$ are fixed randomly according to the transition matrix P of $\mathcal{M}$, then the CFTP of $\mathcal{M}$ halt in finite time with probability 1 and returns a sample according to π.

Proof. Note that, if our algorithm could **cover** all the cases of the last finite steps of the Markov Chain, and could **recover** the sample from all these cases, then its easy to see, this algorithm should be right.

So, here, its quite nature to ask, what means all the cases? Here, a case is a suffix of the chain from $-\infty$ to 0. Note that, CFTP randomly choose a f_{-t} at time $-t$, so if we omit the stop condition, its should cover all the cases. Suppose the CFTP stops at some time T, which means its forms a constant function f and no matter how we extend this suffix, this constant will not change, so we could pack all the suffix that end with this suffix into this suffix, since they have the same sample (while do not change the distribution of sample). Hence, CFTP forms a cover of all the cases.

Since $\lim_{t\to\infty} \Pr[f_0 \circ f_1 \circ \cdots \circ f_t] = 1$, the Markov Chain will be a constant function with probability 1 after infinite steps. Which notes us that CFTP should halt in finite time with probability 1.

Finally, since CFTP simulates a Markov Chain which have been executed for infinite steps, so the distribution of the sample should be exactly π. □

1.8.1 Two confusing bad variations of CFTP

Coupling to the future Its nature for us to ask, "why not just coupling to the future?", since this two samplers look nearly the same.

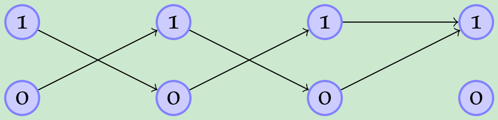
```
Copuling To The Future () begin
    $T \leftarrow 1$
    repeat
        $f \leftarrow f_0 \circ f_1 \circ \cdots \circ f_T$
        $T \leftarrow T + 1$
    until $f$ becomes a constant function;
    return $f(s_1)$
end
```

The problem here is that is sampler does not cover all the cases. First, if we omit the stop condition, then it is easy to see that this sampler covers all the cases. Similarly, this sampler want to pack some of the cases so that we could stop in finite steps. Note that, if f becomes a constant function in time T, then f' will also be a constant function in time $T+1$ and so on. But f and f' has different constant, so we could not pack them together (we should not use the sample of f to represent the sample of f'). So, there is a bias in the result distribution.

Here is a quite simple example to explain this: Suppose we have a Markov chain with state space $[2] = \{1, 2\}$ and transition matrix

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix}$$

Obviously, $\pi = (\pi_1, \pi_2) = (\frac{2}{3}, \frac{1}{3})$. In this example, the *coupling to the future* sampler always returns 1 as its result.



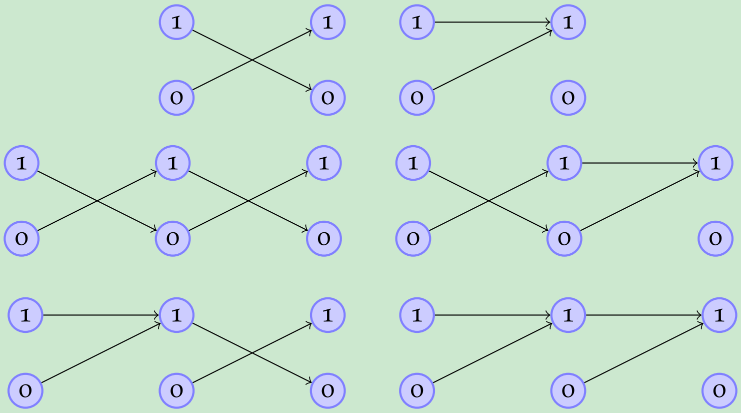
Coupling From The Past with no Memory It is quite nature to implement the CFTP sampler without **reuse**, but unfortunately, this variation of CFTP is wrong. This sampler covers all the cases, but it does not follow the right distribution. The probobility of this sampler returns sample s equals:

$$
\begin{aligned}
&\sum_{T=1}^{\infty} \Pr[f \text{ has constant } s \mid \begin{matrix} f \text{ is a constant function in } T \\ \wedge\ f \text{ is not a constant function in } \{1,\cdots,T-1\} \end{matrix}] \\
&= \sum_{T=1}^{\infty} \Pr[f \text{ has constant } s \mid \text{sampler halts in } T \text{ steps}] \Pr[\text{sampler halts in } T \text{ steps}] \\
&= \sum_{T=1}^{\infty} \frac{\Pr[\text{constant function in } T \text{ steps with constant } s]}{\Pr[\text{function in } T \text{ steps}]} \prod_{t=1}^{T-1} \frac{\Pr[t \text{ steps not constant function}]}{\Pr[\text{function in } t \text{ steps}]}
\end{aligned}
$$

Actually, its hard to find any relationship between this distribution and the correct distribution. When $s = 1$, then using the example above, we have

$$
\begin{aligned}
\Pr[s=1] \geq\ & \Pr[\text{1 step}]\Pr[f \text{ has constant } 1|\text{1 step}] + \Pr[\text{2 steps}]\Pr[f \text{ has constant } 1|\text{2 steps}] \\
&= \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{2}{3} \\
&= \frac{3}{4} > \frac{2}{3}
\end{aligned}
$$

So, its easy to see that this sampler does not catch the right distribution.



2
Topics Related to Probability

2.1 Expectation And Conditional Expectation

Remark 2.1.1. *This content is grabbed from «Markov Chains and Mixing Times, second edition».*

Definition 2.1.1. *random variable*

A random variable X is a measureable function defined on Ω. This means we have $X : \Omega \to \mathbb{R}$.

Definition 2.1.2. *indicator function*

$$\mathbf{1}_A(x) := [x \in A]$$

Actaully, we have defined indicator function before. Here, we only introduce another notation of it.

Definition 2.1.3. *expectation*

For a simple random variable X having the form

$$X = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}$$

,we define

$$\mathbb{E}[X] = \sum_{i=1}^{n} a_i P(A_i)$$

Definition 2.1.4. *conditional expectation*

$$\mathbb{E}[X|A] := \frac{1}{P(A)}\mathbb{E}[X\mathbf{1}_A], \qquad P(A) > 0$$

2.2 Useful Properties for Expectation and Variance

Definition 2.2.1. *Expectation*

For some function $f : \Omega \to \mathbb{R}$, and some distribution π, we could define the expection of f on π as:

$$\mathbb{E}_\pi f = \sum_{x\in\Omega} \pi(x) f(x)$$

.

After we have the concept of expectation, we could define the variance:

Definition 2.2.2. *Variance*

$$\mathrm{Var}_\pi f = \mathbb{E}_\pi (f - \mathbb{E}_\pi f)^2 = \sum_{x\in\Omega} \pi(x)(f(x) - \mathbb{E}_\pi f)^2$$

Actually, besides the definition for the variance, there are many other equal ways for calculating the variance.

Proposition 2.2.1. *Yet Another Way to Calculate Variance*

$$\mathrm{Var}_\pi f = \mathbb{E}_\pi f^2 - (\mathbb{E}_\pi f)^2$$

Proof.

$$\begin{aligned}
\mathrm{Var}_\pi f &= \mathbb{E}_\pi (f - \mathbb{E}_\pi f)^2 \\
&= \sum_{x\in\Omega} \pi(x)(f(x) - \mathbb{E}_\pi f)^2 \\
&= \sum_{x\in\Omega} \pi(x)(f(x)^2 - 2f(x)\mathbb{E}\pi f + (\mathbb{E}_\pi f)^2) \\
&= \sum_{x\in\Omega} \pi(x) f(x)^2 - \sum_{x\in\Omega} \pi(x) 2f(x)\mathbb{E}_\pi f + \sum_{x\in\Omega} (\mathbb{E}_\pi f)^2 \\
&= \mathbb{E}_\pi f^2 - (\mathbb{E}_\pi f)^2 + (\mathbb{E}_\pi f)^2 \\
&= \mathbb{E}_\pi f^2 - (\mathbb{E}_\pi f)^2
\end{aligned}$$

□

Since $\mathrm{Var}_\pi f \geq 0$ from its definition, we have:

Corollary 2.2.1.

$$(\mathbb{E}_\pi f)^2 \leq \mathbb{E}_\pi f^2$$

Proposition 2.2.2. *Yet Another Way to Calculate Variance*

$$\mathrm{Var}_\pi f = \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)(f(x)-f(y))^2$$

Proof.

$$\begin{aligned}
&\frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)(f(x)-f(y))^2\\
&= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)[f(x)^2+f(y)^2-2f(x)f(y)]\\
&= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)^2 + \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)f(y)^2 - \sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)f(y)\\
&= \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)^2 + \frac{1}{2}\sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)^2 - \sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)f(y)
\end{aligned}$$

Here, the mark for x and y does not affect the result of the answer

$$\begin{aligned}
&= \sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)^2 + \sum_{x,y\in\Omega} \pi(x)\pi(y)f(x)f(y)\\
&= \sum_{x,y\in\Omega} [\pi(x)\pi(y)f(x)^2 + \pi(x)\pi(y)f(x)f(y)]\\
&= \sum_{x\in\Omega} \pi(x)f(x)^2 \sum_{y\in\Omega}\pi(y) - \sum_{x\in\Omega}\pi(x)f(x)\sum_{y\in\Omega}\pi(y)f(y)\\
&= \sum_{x\in\Omega} \pi(x)f(x)^2 - (\mathbb{E}_\pi f)^2\\
&= \mathrm{Var}_\pi f
\end{aligned}$$

□

2.3 *Introduction to the Γ Function*

Definition 2.3.1. Γ *function*

$$\Gamma(z) = \int_0^\infty x^{z-1}e^{-x}\mathrm{d}x$$

Though x^z may seems more natural, wiki just defines it like this.

Γ function is a very useful function, it could represent the factorial not in whole nummber but also in real number.

Proposition 2.3.1.

$$\Gamma(z+1) = z\Gamma(z)$$

Proof.

$$
\begin{aligned}
(-x^z e^{-x})' &= x^z e^{-x} + (-z x^{z-1} e^{-x}) \\
\left[-x^z e^{-x}\right]_0^\infty &= \int_0^\infty x^z e^{-x} \mathrm{d}x + \int_0^\infty -z x^{z-1} e^{-x} \mathrm{d}x \\
\int_0^\infty x^z e^{-x} \mathrm{d}x &= \left[-x^z e^{-x}\right]_0^\infty + \int_0^\infty z x^{z-1} e^{-x} \mathrm{d}x \\
\int_0^\infty x^z e^{-x} \mathrm{d}x &= z \int_0^\infty x^{z-1} e^{-x} \mathrm{d}x \\
\Gamma(z+1) &= z\Gamma(z)
\end{aligned}
$$

□

Actually, all the ∞ here means $+\infty$.

Proposition 2.3.2.

$$\Gamma(1) = 1$$

Proof.

$$
\begin{aligned}
\Gamma(1) &= \int_0^\infty x^{1-1} e^{-x} \mathrm{d}x \\
&= \int_0^\infty e^{-x} \mathrm{d}x \\
&= \left[-e^{-x}\right]_0^\infty \\
&= \lim_{x\to\infty} -e^{-x} + e^0 \\
&= 0 + 1 \\
&= 1
\end{aligned}
$$

□

Actually, all the ∞ here means $+\infty$.

Corollary 2.3.1.

$$\Gamma(z) = (z-1)!$$

2.4 *Tricks for Total Variation Distance*

Proposition 2.4.1.

Suppose we have a transition matrix P, and a sequence of distribution π_t, where $\pi_t = \pi_{t-1}P$. Then we have

$$||\pi_{t+1} - \pi_t||_{TV} \leq ||\pi_t - \pi_{t-1}||_{TV}$$

Proof.

$$\begin{aligned}
||\pi_{t+1} - \pi_t||_{TV} &= ||\pi_t P - \pi_{t-1} P||_{TV} \\
&= \frac{1}{2} \max_{||z||_\infty \le 1} (\pi_t - \pi_{t-1}) P z \\
&\le \frac{1}{2} \max_{||w||_\infty \le 1} (\pi_t - \pi_{t-1}) \\
&= ||\pi_t - \pi_{t-1}||_{TV}
\end{aligned}$$

□

Proposition 2.4.2. *Triangle Inequality*

Suppose we have three distribution a, b, c. Then we have

$$||a - c||_{TV} \le ||a - b||_{TV} + ||b - c||_{TV}$$

Proof.

$$\begin{aligned}
||a - c||_{TV} &= \frac{1}{2} \sum_{x \in \Omega} |a(x) - c(x)| \\
&\le \frac{1}{2} \sum_{x \in \Omega} |a(x) - b(x)| + |b(x) - c(x)| \\
&= \frac{1}{2} \sum_{x \in \Omega} |a(x) - b(x)| + \frac{1}{2} \sum_{x \in \Omega} |b(x) - c(x)| \\
&= ||a - b||_{TV} + ||b - c||_{TV}
\end{aligned}$$

□

2.5 *Chebyshev Inequality*

We have encountered with this inequality for many times. We used to it without understanding it. Today, we are going depth, and we want to find out how it works.

Theorem 2.5.1. *Chebyshev Inequality (form 1)*

$$\Pr[|X - \mathbb{E}[X]| \ge \epsilon] \le \mathrm{Var}[X]/\epsilon^2$$

Proof. *(we only prove this in the continuous condition)*

For convenience, let $\mu = \mathbb{E}[X]$ and $\sigma^2 = \mathrm{Var}[X]$. Then by the definition of $\mathrm{Var}[X]$, we have

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{+\infty} (t - \mu)^2 f(t) \mathrm{d}t, \qquad f \text{ is the dense of the probability} \\
&\ge \int_{-\infty}^{\mu - \epsilon} (t - \mu)^2 f(t) \mathrm{d}t + \int_{\mu + \epsilon}^{+\infty} (t - \mu)^2 f(t) \mathrm{d}t
\end{aligned}$$

Then, in the above expression, we have $|t - \mu| \geq \epsilon$. And thus we have $(t-\mu)^2 \geq \epsilon^2$. So

$$\begin{aligned}
\sigma^2 &\geq \int_{-\infty}^{\mu-\epsilon} \epsilon^2 f(t)\mathrm{d}t + \int_{\mu+\epsilon}^{+\infty} \epsilon^2 f(t)\mathrm{d}t \\
&= \epsilon^2 \left(\int_{-\infty}^{\mu-\epsilon} f(t)\mathrm{d}t + \int_{\mu+\epsilon}^{+\infty} f(t)\mathrm{d}t \right) \\
&= \epsilon^2 \Pr\{X \leq \mu - \epsilon \text{ or } X \geq \mu + \epsilon\} \\
&= \epsilon^2 \Pr[|X - \mu| \geq \epsilon]
\end{aligned}$$

□

Theorem 2.5.2. *Chebyshev Inequality (form 2)*

$$\Pr[X - \mathbb{E}[X] \geq \epsilon] \leq \mathrm{Var}[X]/(\mathrm{Var}[X] + \epsilon^2)$$

Proof. *(we only prove this in the continuous condition)*

For convenience, let $\sigma^2 = \mathrm{Var}[X]$ and $\mu = \mathbb{E}[X]$. Then

$$\begin{aligned}
\sigma^2 &= \int_{-\infty}^{\mu+\epsilon} (t-\mu)^2 f(t)\mathrm{d}t + \int_{\mu+\epsilon} (t-\mu)^2 f(t)\mathrm{d}t \\
&\geq \int_{-\infty}^{\mu} (t-\mu)^2 f(t)\mathrm{d}t + \int_{\mu+\epsilon}^{+\infty} \epsilon^2 f(t)\mathrm{d}t
\end{aligned}$$

Since $\int_{-\infty}^{\mu} f(t)\mathrm{d}t = 1/2$, we have

$$\int_{-\infty}^{\mu} (t-\mu)^2 f(t)\mathrm{d}t = \frac{1}{2}\sigma^2$$

So

$$\begin{aligned}
\sigma^2 &\geq \frac{1}{2}\sigma^2 + \int_{\mu+\epsilon}^{+\infty} \epsilon^2 f(t)\mathrm{d}t \\
&= \sigma^2 \Pr[X \geq \mu] + \epsilon^2 \Pr[X \geq \mu + \epsilon] \\
&\geq \sigma^2 \Pr[X \geq \mu + \epsilon] + \epsilon^2 \Pr[X \geq \mu + \epsilon] \\
&= (\sigma^2 + \epsilon^2) \Pr[X - \mu \geq \epsilon]
\end{aligned}$$

□

2.6 *Proportion Inequality Between Two Distribution*

Actually, I do not know if this should be called proportion inequality. I Just find that it is derived by the proportion of two distributions and it is really interesting. I stuck at the first time when I met it.

Lemma 2.6.1. *Proportion Inequality*

For two distribution π and μ which have the same ground set Ω, let

$$Z = \{x \in \Omega | \pi(x)/\mu(x) \geq \varrho\}, \quad 0 \leq \varrho \leq 1$$

Then

$$\pi(Z) \geq 1 - \varrho$$

Proof. Consider $\overline{Z}$. It is clear that $\pi(\overline{Z}) < \varrho$, since

$$\pi(\overline{Z}) \leq \sum_{x \in \Omega} \varrho\mu(x) < \varrho$$

Thus, we have

$$\pi(Z) \geq 1 - \varrho \qquad \square$$

Lemma 2.6.2. *Total Variation Distance of a Proportion Set*

For two distribution π and μ which have the same ground set Ω, and

$$||\pi - \mu||_{TV} \leq \varepsilon, \quad 0 \leq \varepsilon \leq 1$$

. Let

$$Z = \{x \in \Omega | \pi(x)/\mu(x) \geq \varrho\}, \quad 0 \leq \varrho \leq 1$$

Then

$$\pi(Z) \geq \frac{1 - \varrho - \varepsilon\varrho}{1 - \varrho}$$

and

$$\mu(Z) \geq \frac{1 - \varrho - \varepsilon}{1 - \varrho}$$

Remark 2.6.1. *Actually if we conbine the above two lemma, we have:*

$$\pi(Z) \geq \max\{1 - \varrho, \frac{1 - \varrho - \varepsilon\varrho}{1 - \varrho}\}$$

Proof. First we have:

$$\begin{aligned} \sum_{x \in \overline{Z}} \mu(x) - \pi(x) &> \sum_{x \in \overline{Z}} (\frac{1}{\varrho} - 1)\pi(x) \\ \varepsilon &> (\frac{1}{\varrho} - 1)\pi(\overline{Z}) \\ \pi(\overline{Z}) &< \frac{\varrho\varepsilon}{1 - \varrho} \\ \pi(Z) &\geq \frac{1 - \varrho - \varrho\varepsilon}{1 - \varrho} \end{aligned}$$

Second we have:

$$\begin{aligned}
\sum_{x \in \overline{Z}} \mu(x) - \pi(x) &> \sum_{x \in \overline{Z}} (1-\varrho)\mu(x) \\
\varepsilon &> (1-\varrho)\mu(\overline{Z}) \\
\mu(\overline{Z}) &< \frac{\varepsilon}{1-\varrho} \\
\mu(Z) &\geq \frac{1-\varrho-\varepsilon}{1-\varrho} \qquad \square
\end{aligned}$$

3
Topics Related to Geometry

3.1 Radon's Theorem

Theorem 3.1.1. *Radon's Theorem*

Any set of $d+2$ points in $\mathbb{R}^d$ can be partitioned into two disjoint sets whose convex hulls intersect. A point in the intersection of these convex hulls is called a Radon point of the set.

Proof. Suppose there is a set $X = \{x_1, x_2, \cdots, x_{d+2}\} \subset \mathbb{R}^d$ of $d+2$ points in d-dimentinal space. Then there exists a set of multipliers $a_1, a_2, \cdots, a_{d+2}$, not all of which are zero, solving the system of linear equations:

$$\sum_{i=1}^{d+2} a_i x_i = 0, \qquad \sum_{i=1}^{d+2} a_i = 0$$

Since there are only $d+1$ equations but $d+2$ variables, there are some non-zero solutions for this system. Suppose we have already solved this system and have a non-zero answer for a's. Let $I = \{x_i | a_i > 0\}$ and $J = X \setminus I$. Then we find that the convex hull of I and J has common point, which means there convex hulls intersect. Clearly, I and J must intersect, because they both contain the point:

$$p = \sum_{i \in I} \frac{a_i}{A} x_i = \sum_{j \in J} \frac{-a_j}{A} X_j$$

where

$$A = \sum_{i \in I} a_i = -\sum_{j \in J} a_j$$

Since the right hand side of p represent a convex combination of the points in I and the points in J, p belongs to both convex hulls. This also gives us a efficient way to construct a Radon point. □

3.2 Helly's Theorem

Theorem 3.2.1. *Helly's Theorem*

Let $X_1, X_2, \cdots, X_n$ be a finite collection of convex subsets of $\mathbb{R}^d$, with $n > d$. If the intersection of every $d+1$ of these sets is nonempty, then the whole collection has a nonempty intersection, that is:

$$\bigcap_{j=1}^{n} X_j \neq \emptyset$$

Proof. For convenient, we use $\mathrm{Cov}(A)$ to represent the convex of point set A. We prove this by induction.

Base Case ($n = d+2$): Then for every $j = 1, \cdots, n$ there is a point x_j that is in the common intersection of all X_i with the possible exception of X_j. Now we apply Radon's theorem to the set $A = \{x_1, \cdots, x_n\}$, which furnishes us with disjoint subsets A_1, A_2 of A such that

$$\text{Cov}(A_1) \cap \text{Cov}(A_2) \neq \emptyset$$

Suppose that p is a point in the intersection of these two convex hulls. We claim that

$$p \in \bigcap_{i=1}^{n} X_j$$

Indeed for any $j \in \{1, \cdots, n\}$, we want to show that $p \in X_j$:

1. If $x_j \in X_j$ and $x_j \in \text{Cov}(A_1)$, then we have $A_1 \subseteq X_j$. So $p \in \text{Cov}(A_1) \subseteq X_j$.
2. If $x_j \notin X_j$ and $x_j \in \text{Cov}(A_1)$, then we have $x_j \notin \text{Cov}(A_2)$ and $\text{Cov}(A_2) \subseteq X_j$. So $p \in \text{Cov}(A_2) \subseteq X_j$.

Above, we have assumed that the points $x_1, \cdots, x_n$ are all distinct. If this is not the case, say $x_i = x_k$ for some $i \neq k$, then x_i is in every one of the sets X_j, and again we conclude that the intersection is nonempty. This completes the proof in the case $n = d+2$.

Inductive Step ($n > d+2$) Assume the theorem is true in the case $n-1$. Note that we have n sets $X_1, X_2, \cdots, X_n$ whre $n > d+2$. Actually, for every $d+2$ of these sets, we could apply the statement for the base case for them. Which turns out that the intersection for every $d+2$ sets are not empty. Then if we let $X_{n-1} \leftarrow X_{n-1} \cap X_n$, we could get to a $n-1$ case where the intersection of every $d+1$ sets is not empty. So by our assumption, the intersection of these $n-1$ sets is not empty and thus

$$\bigcap_{i=1}^{n} X_i \neq \emptyset$$

□

3.3 *Brunn-Minkowski Theorem*

The material of this topic is mainly found in a course named An Algorithmist's Toolkit at **MIT OCW**. They are really good at explaing this.

Definition 3.3.1. *The Notation of Minkowski Sum*

Let A and B be sets of points and λ be a real number. A point p is represented by the vector pointing from 0 to p. Then the Minkowshi Sum notation could be defined as below:

$$A + B = \{a + b | a \in A, b \in B\}$$
$$\lambda A = \{\lambda a | a \in A\}$$

Note that this definition has some good properties only when A and B are convex.

Fist, we provide some trivial facts about Vol_n and Minkowshi Sum.

Fact 3.3.1.

$$\text{Vol}_n(A+B) \geq \max\{\text{Vol}_n(A), \text{Vol}_n(B)\}$$

Proof. First, if $A = \emptyset$ or $B = \emptyset$, then this fact is true. Else there exists $a \in A$, and clearly we have $\{a\} + B \subseteq A + B$. Hence,

$$\text{Vol}_n(B) \leq \text{Vol}_n(A+B)$$

. Similarly,

$$\text{Vol}_n(A) \leq \text{Vol}_n(A+B) \qquad \square$$

.

Fact 3.3.2.

$$\text{Vol}_n(A+B) \geq \text{Vol}_n(A) + \text{Vol}_n(B)$$

The bound given by this fact is loose, since we have $\text{Vol}_n(2A) = 2^n\text{Vol}_n(A)$

Proof. Assume A and B are all positive. Then we could choose two points $a \in A$ and $b \in B$ with both has the maximum value in direction e_1 (e.g. x direction). Then its clear that $\{a\} + B$ is disjoint with $\{b\} + A$. So we have:

$$\{a\} + B + \{b\} + A \subseteq A + B$$
$$\text{Vol}_n(\{a\} + B) + \text{Vol}_n(\{b\} + A) \leq \text{Vol}_n(A+B) \qquad \square$$

Fact 3.3.3.

Move A and B will not change the volume of A, B, or $A+B$.

Proof. Suppose we want to move A by a vector a and move B by a vector b. Then we have:

$$A + a = \{x + a | x \in A\}$$
$$B + b = \{x + b | x \in B\}$$
$$(A+a) + (B+b) = \{x + (a+b) | x \in A + B\}$$

Clearly, these three are bijection betweent A and $A+a$, B and $B+b$, $A+B$ and $A+B+a+b$. So, the move of these point sets will not change their volume. $\square$

Now we are going to get a tight bound on this:

Lemma 3.3.1. *Burnn-Minkowshi Theorem on Box*

Let A and B be box in $\mathbb{R}^n$. Then:

$$\text{Vol}_n(A+B)^{1/n} \geq \text{Vol}_n(A)^{1/n} + \text{Vol}_n(B)^{1/n}$$

Proof. Let A have sides of length $a_1, a_2, \cdots, a_n$, B have sides of length $b_1, b_2, \cdots, b_n$. Then the Minkowshi Sum $A+B$ has sides of length $a_1+b_1, a_2+b_2, \cdots, a_n+b_n$. Then:

$$\begin{aligned}
\frac{\text{Vol}_n(A)^{1/n} + \text{Vol}_n(B)^{1/n}}{\text{Vol}_n(A+B)^{1/n}} &= \frac{(\prod_{i=1}^n a_i)^{1/n} + (\prod_{i=1}^n b_i)^{1/n}}{(\prod_{i=1}^n (a_i+b_i))^{1/n}} \\
&= \prod_{i=1}^n \left(\frac{a_i}{a_i+b_i}\right)^{1/n} + \prod_{i=1}^n \left(\frac{b_i}{a_i+b_i}\right)^{1/n} \\
&\leq \frac{1}{n}\sum_{i=1}^n \frac{a_i}{a_i+b_i} + \frac{1}{n}\sum_{i=1}^n \frac{b_i}{a_i+b_i} \\
&\leq 1
\end{aligned}$$ □

Now, we want to generize the result to any convexi point sets A, and B.

Theorem 3.3.1. *Brunn-Minkowski Theorem*

Let A and B be convex point sets (or measurable point sets). Then:

$$\text{Vol}_n(A+B)^{1/n} \geq \text{Vol}_n(A)^{1/n} + \text{Vol}_n(B)^{1/n}$$

Proof. We prove this by induction. We pick the base case to be two boxes A and B, which is proved in the lemma above.

In the inductive case, A and B are made up by finite number of disjoint boxes. As the number of boxes converge to ∞, we could simulate any convex point set.

Define the following subsets of $\mathbb{R}^n$:

$$\begin{aligned}
A^+ &= A \cap \{x \in \mathbb{R}^n | x_n \geq 0\}, \quad & A^- &= A \cap \{x \in \mathbb{R}^n | x_n \leq 0\} \\
B^+ &= B \cap \{x \in \mathbb{R}^n | x_n \geq 0\}, \quad & B^- &= B \cap \{x \in \mathbb{R}^n | x_n \leq 0\}
\end{aligned}$$

Move A and B such that the following conditions holds:

1. A has some pair of boxes separated by the hyperplane $\{x \in \mathbb{R}^n | x_1 = 0\}$. i.e. there exists a box that lies completely in the halfspace $\{x \in \mathbb{R}^n | x_1 \geq 0\}$ and there is some other box that lies in its complement half-space.

2. It holds that
$$\frac{\text{Vol}_n(A^+)}{\text{Vol}_n(A)} = \frac{\text{Vol}_n(B^+)}{\text{Vol}_n(B)}$$

Actually, we could split A into two equal volume parts and split each part recurrently. We could do the same operation on B. This operation could stop when the volume of each part is small enough. In this way, the above two requirement could always be achieved.

Then because we achieved condition 1, we could find that A^+ and A^- are strictly subset of A. So $A^+ \cup B^+$ and $A^- \cup B^-$ have fewer boxes than $A \cup B$ and the inductive hypothesis is ture on them. Moreover, $A^+ \cup B^+$ and $A^- and B^-$ are disjoint because they have different sign of the x_1 coordinate. Hence we have:

$$\begin{aligned}
\text{Vol}_n(A+B) &\geq \text{Vol}_n(A^+ + B^+) + \text{Vol}_n(A^- + B^-) \\
&\geq (\text{Vol}_n(A^+)^{1/n} + \text{Vol}_n(B^+)^{1/n})^n + (\text{Vol}_n(A^-)^{1/n} + \text{Vol}_n(B^-)^{1/n})^n, \quad \text{by inductive hypothesis} \\
&= \text{Vol}_n(A^+)(1 + \frac{\text{Vol}_n(B^+)^{1/n}}{\text{Vol}_n(A^+)^{1/n}})^n + \text{Vol}_n(A^-)(1 + \frac{\text{Vol}_n(B^-)^{1/n}}{\text{Vol}_n(A^-)^{1/n}})^n \\
&= \text{Vol}_n(A^+)(1 + \frac{\text{Vol}_n(B)^{1/n}}{\text{Vol}_n(A)^{1/n}})^n + \text{Vol}_n(A^-)(1 + \frac{\text{Vol}_n(B)^{1/n}}{\text{Vol}_n(A)^{1/n}})^n, \quad \text{by condition 2} \\
&= (\text{Vol}_n(A^+) + \text{Vol}_n(A^-))(1 + \frac{\text{Vol}_n(B)^{1/n}}{\text{Vol}_n(A)^{1/n}})^n \\
&= \text{Vol}_n(A)(1 + \frac{\text{Vol}_n(B)^{1/n}}{\text{Vol}_n(A)^{1/n}})^n \\
&= (\text{Vol}_n(A)^{1/n} + \text{Vol}_n(B)^{1/n})^n
\end{aligned}$$

□

Here we have an equivalent for Brunn-Minkowshi Theorem:

Corollary 3.3.1.

$$\text{Vol}_n(\lambda A + (1-\lambda)B)^{1/n} \geq \lambda \text{Vol}_n(A)^{1/n} + (1-\lambda)\text{Vol}_n(B)^{1/n}$$

Proof.

$$\begin{aligned}
\text{Vol}_n(\lambda A + (1-\lambda)B)^{1/n} &\geq \text{Vol}_n(\lambda A)^{1/n} + \text{Vol}_n((1-\lambda)B)^{1/n} \\
&= (\lambda^n \text{Vol}_n(A))^{1/n} + ((1-\lambda)^n \text{Vol}_n((1-\lambda)B))^{1/n} \\
&= \lambda \text{Vol}_n(A)^{1/n} + (1-\lambda)\text{Vol}_n((1-\lambda)B)^{1/n}
\end{aligned}$$

□

4
Topics in Measure Theory

4.1 Basic Notations and Ideas

4.1.1 Non-measurealbe Sets

Measure is very important since the integrating and probability are built on top of it. At the begining, the idea is very simple:

The concept for the measure of length grows up naturally. For a set $(a,b] \in \mathbb{R}$, the length of it is $b-a$. As the interval $(a,b]$ could be seen as a subset of $\mathbb{R}$, we might ask, could we define a function $f : \mathcal{P}(\mathbb{R}) \to \mathbb{R}$ to give a measure for all the subset of $\mathbb{R}$?

At first, some people start to find such a function f on $\mathcal{P}(\mathbb{R})$. Ideally, if we want to define a function λ like this, we'd like it satisfies some properties:

- (0) $\lambda : \mathcal{P}(\mathbb{R}) \to \mathbb{R}_+ \cup \{+\infty\}$
- (1) $\lambda((a,b]) = b-a$, since we want it to remain its nature meaning.
- (2) $\lambda(A+x) = \lambda(A)$, since the shift should not change the volume.
- (3) $\lambda(\cup_{j\geq 1} A_j) = \sum_{j\geq 1} \lambda(A_j)$ for coutable disjoint A_j. Since we need to add the volume for calculation.

But soon, they find it impossible, as we could prove this function could not exists. Otherwise, there must be a contradiction.

Theorem 4.1.1.

The function that satisfies the requirements above does not exist.

Proof. For $x, y \in \mathbb{R}$, let $x \sim y$ if $y - x \in \mathbb{Q}$. Let $[x] = \{y \in \mathbb{R} | y - x \in \mathbb{Q}\}$ be a equivalent class. Let $\Lambda = \mathbb{R}| \sim$, which means the set of equivalent classes. Now, we construct a set $\Omega \in \mathbb{R}$ be the set which contains one and only one points from each equivalent class. Thus we could assume that $\Omega \subseteq (0,1)$.

Λ is uncountable, since $\mathbb{R}$ is uncountable, and any $[x] \in \Lambda$ is countable. We may use α, β to represent the point in Λ.

Claim 4.1.1. *For a set $\Omega \subseteq \mathbb{R}$, and $p, q \in \mathbb{R}$. Either we have $\Omega + p = \Omega + q$, eigher we have $\Omega + p$ disjoints with $\Omega + q$*

Proof. First, lets assume $(\Omega + p) \cap (\Omega + q) \neq \emptyset$. Then for some x in it, we have

$$\begin{aligned} x &= \alpha + p, \quad \alpha \in \Omega \\ &= \beta + q, \quad \beta \in \Omega \end{aligned}$$

So we have $\alpha - \beta = q - p$. From the definition of Ω, we know that $p = q$ and $\alpha = \beta$. So we have

$$q \neq p \Rightarrow (\Omega + q) \cap (\Omega + p) = \emptyset$$

. □

By this claim, we could define a set:

$$S = \sum_{\substack{q \in \mathbb{Q} \\ -1 < q < 1}} (\Omega + q)$$

Its easy to see that for all these q, $\Omega + q \subseteq (-1, 2)$. So we have $S \subseteq (-1, 2)$ and $\lambda(S) \leq \lambda((-1, 2)) = 3$. So by property (3), we have

$$\sum_{\substack{q \in \mathbb{Q} \\ -1 < q < 1}} \lambda(\Omega + q) \leq 3$$

And by property (2), we have $\lambda(\Omega + q) = \lambda(\Omega) = 0$. By observing this, we have $\lambda(S) = 0$.

Here, since we have infinity qs, $\lambda(\Omega + q) = 0$. Otherwise, the sum of them could not less than 3.

Claim 4.1.2. $(0, 1) \subseteq S$

Proof. Suppose we have $x \in (0, 1)$ and $\alpha \in [x] \cap \Omega$, then we have $\alpha \in (0, 1)$ (by definition of Ω). Since α belongs to the equivalent class of $[x]$, $\alpha - x = q \in \mathbb{Q}$. Its easy to see that $-1 < q < 1$. Which means $x = \alpha + q$ and thus $x \in \Omega + q$ for some $q \in (0, 1)$. Which is exactly in S. So we have $(0, 1) \subseteq S$. □

So far, we have find a contradiction that there is no such function which could satisfies all the properties. □

What can we do now? Maybe we could relax some of these properties. Maybe we could remove some this conditions. Maybe we could accept the function which is not defined on all the subsets of $\mathbb{R}$. This means there are some kinds of subsets of $\mathcal{P}$ which we could not set a measure on it, these sets are called non-measurealbe sets.

4.1.2 Classes of subsets, and set functions

Definition 4.1.1. *semi-algebra*

$\mathcal{S} \subseteq \mathcal{P}(\Omega)$, is a semi-algebra if:

1. $\Omega \in \mathcal{S}$.
2. $A, B \in \mathcal{S} \Rightarrow A \cap B \in \mathcal{S}$.

   (*This means it is closed under finite intersections*)
3. $\forall A \in \mathcal{S} \Rightarrow \exists E_1, E_2, \cdots, E_n \in \mathcal{S}$ such that $A^c = \cup_{j=1}^{n} E_j$.

Definition 4.1.2. *algebra*

$\mathcal{A} \subseteq \mathcal{P}(\Omega)$, is an algebra if:

1. $\Omega \in \mathcal{A}$.
2. $A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}$.
3. $A \in \mathcal{A} \Rightarrow A^c \in \mathcal{A}$.

   (*Here is the difference between semi-algebra and algebra*)

Definition 4.1.3. *σ-algebra*

$\mathcal{F} \subseteq \mathcal{P}(\Omega)$, is an algebra if:

1. $\Omega \in \mathcal{F}$.
2. $A_j \in \mathcal{F}$ for $j \geq 1 \Rightarrow \cap_{j \geq 1} A_j \in \mathcal{F}$.

   (*closed under countable intersections*)

Remark 4.1.1. *σ-algebra $\Rightarrow$ algebra $\Rightarrow$ seim-algebra. σ-algebra asks for more constraint than algebra, and thus it is bigger than an algebra (this means you could find some algebra in σ-algebra or σ-algebra is itself an algebra).*

Observation 4.1.1.

If $\mathcal{A}_\alpha \subseteq \mathcal{P}(\Omega)$ is (σ)-algebra for some $\alpha \in I$. Then we have $\mathcal{A} = \cap_{\alpha \in I} \mathcal{A}_\alpha$ is a (σ)-algebra.

we can use these fact to introduce the notion of the algebra generated by the classes of sets.

Definition 4.1.4.

The algebra generated by class $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ is denoted by $\mathcal{A}(\mathcal{C})$. And $\mathcal{A}(\mathcal{C})$ satisfies:

- $\mathcal{C} \subseteq \mathcal{A}(\mathcal{C})$, $\mathcal{A}(\mathcal{C})$ is algebra.
-
$$\begin{cases} \mathcal{C} \subseteq \mathcal{B} \\ \mathcal{B} \text{ is algebra} \end{cases} \Rightarrow \mathcal{A}(\mathcal{C}) \subseteq \mathcal{B}$$

Let $\mathcal{A}_\alpha$ be all the algebra that contains $\mathcal{C}$. Then it is easy to verify that

$$\mathcal{A}(\mathcal{C}) = \bigcap_\alpha \mathcal{A}_\alpha$$

.

Lemma 4.1.1.

$\mathcal{S} \subseteq \mathcal{P}(\Omega)$ is a semi-algebra, $\mathcal{A}$ is the algebra generated by $\mathcal{S}$. Then

$$A \in \mathcal{A} \Leftrightarrow \substack{\exists E_j, i \leq j \leq n, E_j \in \mathcal{S} \\ A = \sum_{j=1}^n E_j}$$

Proof. *(⇐):* This is obvious from the definition of $\mathcal{A}$.

(⇒): Define a new class

$$\mathcal{B} = \{\sum_{j=1}^n F_j, F_j \in \mathcal{S}\}$$

. We will prove that:

$$\begin{cases} (1)\ \mathcal{B} \text{ is an algebra} \\ (2)\ \mathcal{S} \subseteq \mathcal{B} \end{cases} \Rightarrow \mathcal{A} \subseteq \mathcal{B}$$

Note that (2) is trivial since every element in $\mathcal{S}$ could be written in the form of $\sum_{j=1}^n F_j$.

Now we prove the (1).

1. $\Omega \in \mathcal{B}$. (this is trivial, since $\Omega \in \mathcal{S} \subseteq \mathcal{B}$).

2. $A, B \in \mathcal{B} \Rightarrow A \cap B \in \mathcal{B}$. (easy to verify)

3. $A \in \mathcal{B} \Rightarrow A^c \in \mathcal{B}$.
   Let $A \in \mathcal{B}$ and $A = \sum_{j=1}^n E_j, E_j \in \mathcal{S}$. Then $A^c = (\sum_{j=1}^n E_j)^c = E_1^c \cap E_2^c \cap \cdots \cap E_n^c$. And $E_i^c = \sum_{k_i=1}^{l_i} F_{i,k_i}, F_{i,j} \in \mathcal{S}$ (This is by the

definition of semi-algebra). So,

$$\begin{aligned} A^c &= (\sum_{k_1=1}^{l_1} F_{1,k_1}) \cap (\sum_{k_2=1}^{l_2} F_{2,k_2}) \cap \cdots \cap (\sum_{k_n=1}^{l_n} F_{n,k_n}) \\ &= \sum_{k_1=1}^{l_1} \sum_{k_2=1}^{l_2} \cdots \sum_{k_n=1}^{l_n} (F_{1,k_1} \cap F_{2,k_2} \cap \cdots \cap F_{n,k_n}) \end{aligned}$$

So, $A^c \in \mathcal{B}$.

Since $\mathcal{B}$ is an algebra, and it contains $\mathcal{S}$, so $\mathcal{A} \subseteq S$. □

Now we investigate functions defined on these sets.

Definition 4.1.5. *addtive function*

For some $\mathcal{C} \subseteq \mathcal{P}(\Omega)$ where $\emptyset \in \mathcal{C}$. we could define a function $\mu : \mathcal{C} \to \mathbb{R}_+ \cup \{+\infty\}$. Then μ is additive when:

1. $\mu(\emptyset) = 0$.
2. $\substack{E_1, E_2, \cdots E_n \in \mathcal{C} \\ E = \sum_{j=1}^n E_j j} \Rightarrow \mu(E) = \sum_{j=1}^n \mu(E_j)$
   (defined on finite disjoint union)

Observation 4.1.2.

If $E \subseteq F \subseteq \mathcal{C}C$ and $\mu : \mathcal{C} \to \mathbb{R}_+ \cup \{+\infty\}$ is addtive, then

1. $\mu(E) = +\infty \Rightarrow \mu(F) = +\infty$.
2. $\mu(E) < +\infty \Rightarrow \mu(F) = \mu(E) + \mu(F \setminus E)$.

So, $\mu(E) \le \mu(F)$ and when $\mu(E)$ is finite, then

$$\mu(F \setminus E) = \mu(F) - \mu(E)$$

Definition 4.1.6. *σ-addtive function*

For some $\mathcal{C} \subseteq \mathcal{P}(\Omega)$, where $\emptyset \in \mathcal{C}$. We could define a function $\mu : \mathcal{C} \to \mathbb{R}_+ \cup \{+\infty\}$. Then μ is σ-additive when:

1. $\mu(\emptyset) = 0$.
2. $\substack{E_j \in \mathcal{C}, j \ge 1 \\ E_j \cap E_k = \emptyset \\ E = \sum_{j \ge 1} E_j} \Rightarrow \mu(E) = \sum_{j \ge 1} \mu(E_j)$
   (defined on countable disjoint union)

From the definition above, we could define what is a continuous measure (note that a measure is actually a function)

Definition 4.1.7. *continuous from below*

Suppose we have

$$\substack{\mathcal{C}\subseteq\mathcal{P}(\Omega)\\ \mu:\mathcal{C}\to\mathbb{R}_+\cup\{+\infty\}\\ E\in\mathcal{C}}$$

, then when we say μ continuous from below at E, we means:

$$\forall\{E_n\}_{n\geq1,\substack{E_n\in\mathcal{C}\\ E_n\uparrow E}} \Rightarrow \mu(E_n)\uparrow\mu(E)$$

Here, when we say $E_n \uparrow E$, we means $E_n \subseteq E_{n+1}$ and $\bigcup_{n\geq1} E_n = E$.

Definition 4.1.8. *continuous from above*

Suppose we have

$$\substack{\mathcal{C}\subseteq\mathcal{P}(\Omega)\\ \mu:\mathcal{C}\to\mathbb{R}_+\cup\{+\infty\}\\ E\in\mathcal{C}}$$

, then when we say μ continuous from above at E, we means:

$$\forall\{E_n\}_{n\geq1,\substack{E_n\in\mathcal{C}\\ E_n\downarrow E\\ \exists n_0,\mu(n_0)\leq+\infty}} \Rightarrow \mu(E_n)\downarrow\mu(E)$$

Here, when we say $E_n \downarrow E$, we means $E_n \supseteq E_{n+1}$ and $\bigcap_{n\geq1} E_n = E$.

(Note that here, there must exists some n_0 for $\mu(n_0) \leq +\infty$. In my mind, this is because, $+\infty$ is outside of $\mathbb{R}$, but we could only perform analysis inside $\mathbb{R}$. So, we need some tools to pull us back.)

Definition 4.1.9.

If a function continuous from both above and below, then we call it continuous.

Lemma 4.1.2.

Suppose we have

$$\substack{\mathcal{A}\subseteq\mathcal{P}(\Omega),\ \textit{algebra}\\ \mu:\mathcal{A}\to\mathbb{R}_+\cup\{+\infty\},\ \textit{additive}}$$

, then we have:

1. μ *is* σ*-additive* $\Rightarrow$ μ *is continuous at* $E, \forall E \in \mathcal{A}$.
2. μ *is continuous from below* $\Rightarrow$ μ *is* σ*-additive.*
3. $\substack{\mu \text{ is continuous from above at } \emptyset\\ \mu \text{ is finite } (\mu(\Omega)<+\infty)} \Rightarrow \mu$ *is* σ*-additive.*

Proof. *(1) (μ is σ-additive $\Rightarrow$ μ is continuous at $E, \forall E \in \mathcal{A}$ from below):*

Since we have $E_n \uparrow E$, then we define $F_n = E_n \setminus E_{n-1}$. Its clear that F_k is disjoint and $\cup_{n\geq1}E_n = \cup_{k\geq1}F_k, \sum_{k=1}^n F_k = E_n$. So, using the

σ-additive of μ, we have

$$\begin{aligned} \mu(E) = \sum_{k\geq 1} F_k &= \lim_{n\to\infty} \sum_{k=1}^{n} \mu(F_k) \\ &= \lim_{n} \mu(\sum_{k=1}^{n} F_k), \qquad \text{using additive} \\ &= \lim \mu(E_n) \end{aligned}$$

(1) (μ is σ-additive $\Rightarrow$ μ is continuous at $E, \forall E \in \mathcal{A}$ from <u>above</u>):

Recall what we have:

$$\begin{cases} E \in \mathcal{A} \\ E_n \in \mathcal{A}, E_n \downarrow E \\ \mu(E_{n_0}) < +\infty \end{cases}$$

Lets define $G_k = E_{n_0} \setminus E_{n_0+k}$. Since $E_{n_0+1}, E_{n_0+2}, \cdots$ is getting smaller and smalller, we have $G_k \uparrow E_{n_0} \setminus E$. Thus, as we have proved above, $\mu(G_k) \uparrow \mu(E_{n_0} \setminus E)$. Since $\mu(E_{n_0})$ is finite, we have

$$\begin{aligned} \mu(E_{n_0} \setminus E) &= \lim_{k} \mu(E_{n_0} \setminus E_{n_0+k}) \\ \mu(E_{n_0}) - \mu(E) &= \lim_{k} (\mu(E_{n_0}) - \mu(E_{n_0+k})) \\ \mu(E) &= \lim_{k} E_{n_0+k} \\ \mu(E) &= \lim E_n \end{aligned}$$

(2) (μ is continuous from below $\Rightarrow$ μ is σ-additive):

Suppose we have $E = \sum_{k\geq 1} E_k (E, E_k \in \mathcal{A})$, we want to prove that $\mu(E) = \sum_{k\geq 1} \mu(E_k)$. Let $F_n = \sum_{k=1}^{n} E_k$, then $F_n \uparrow E$. Since μ is continuous from below, we have $\mu(F_n) \uparrow \mu(E)$. And by additive $\sum_{k=1}^{n} \mu(E_k) \uparrow \mu(E)$. If we write this limit in infinity sum form we have

$$\begin{aligned} \lim_{n\to+\infty} \sum_{k=1}^{n} \mu(E_k) &= \mu(E) \\ \lim_{k\geq 1} \mu(E_k) &= \mu(E) \end{aligned}$$

Here, the notation

$$\sum_{k\geq 1}$$

could be seen as an abbreviate of

$$\lim_{n\to+\infty} \sum_{k=1}^{n}$$

(this is sometimes called infinity sum)

(3) ($\begin{matrix} \mu \textit{ is continuous from above at } \emptyset \\ \mu \textit{ is finite, which means } \mu(\Omega) < +\infty \end{matrix}$ $\Rightarrow$ μ is σ-additive):

Suppose we have $E = \sum_{k\geq 1} E_k, (E, E_k \in \mathcal{A})$, we want to show that $\mu(E) = \sum_{k\geq 1} \mu(E_k)$. Let $F_n = \sum_{k\geq n} E_k = E \setminus \sum_{j=1}^{n-1} E_j \in \mathcal{A}$. So we have $F_n \downarrow \emptyset$.

Since we have $\substack{F_n \in \mathcal{A} \\ F_n \downarrow \emptyset \\ \mu(F_n) < +\infty}$, by the definition of continuous from above at $\emptyset$, we have $\mu(F_n) \downarrow \mu(\emptyset) = 0$. Thus, we have

$$\begin{aligned}
\mu(E) &= \mu(\sum_{k=1}^{n} E_k \cup \sum_{k>n} E_k) \\
&= \sum_{k=1}^{n} \mu(E_k) + \mu(F_{n+1}), \quad \text{by additive} \\
&= \lim_{n\to+\infty} (\sum_{k=1}^{n} \mu(E_k) + \mu(F_{n+1})), \quad \text{since the above equality holds for all } n \\
&= \lim_{n\to+\infty} (\sum_{k=1}^{n} \mu(E_k)) + 0 \\
&= \sum_{k\geq 1} \mu(E_k)
\end{aligned}$$

□

Having these definitions and properties, we could start our extension.

Theorem 4.1.2. *extension semi-algebra to algebra with (σ)-additive reserved*

If $\mathcal{S} \subseteq \mathcal{P}(\Omega)$
$\mu : \mathcal{S} \to \mathbb{R}_+ \cup \{+\infty\}$, (σ)-additive.
Then $\exists \nu : \mathcal{A}(\mathcal{S}) \to \mathbb{R}_+ \cup \{+\infty\}$, such that:

1. *ν is (σ)-additive.*
2. *$\nu(A) = \mu(A), \forall A \in \mathcal{S}$.*
3. *ν is unique. Which means:*
   $$\substack{\nu_1,\nu_2:\mathcal{A}(\mathcal{S})\to\mathbb{R}_+\cup\{+\infty\} \\ \nu_1(A)=\nu_2(A),\forall A\in\mathcal{S}} \Rightarrow \nu_1(E) = \nu_2(E), \forall E \in \mathcal{A}(\mathcal{S})$$

Proof. Since any $A \in \mathcal{A}(\mathcal{S})$ could be written in the form

$$A = \sum_{j=1}^{n} E_j, \quad E_j \in \mathcal{S}$$

(note that this property only exists between semi-algebra and algebra) So, to remains additive, we could define

$$\nu(A) \overset{add}{=\!=} \sum_{j=1}^{n} \nu(E_j) \overset{ext}{=\!=} \sum_{j=1}^{n} \mu(E_j)$$

(1) (ν is well-defined):

Suppose A has two representations

$$A = \sum_{j=1}^{n} E_j = \sum_{k=1}^{m} F_k, \quad E_j, F_k \in \mathcal{S}$$

By additive, we have

$$\begin{aligned}\sum_{j=1}^{n} E_j &= \sum_{j=1}^{n} E_j \cap A \\ &= \sum_{j=1}^{n} E_j \cap (\sum_{k=1}^{m} F_k) \\ &= \sum_{j=1}^{n}\sum_{k=1}^{m} E_j \cap F_k \\ \sum_{j=1}^{n} \mu(E_j) &= \sum_{j=1}^{n}\sum_{k=1}^{m} \mu(E_j \cap F_k)\end{aligned}$$

Similarly,

$$\sum_{k=1}^{m} \mu(F_k) = \sum_{j=1}^{n}\sum_{k=1}^{m} \mu(E_j \cap F_k)$$

(2) (ν is additive):

Say we have

$$\begin{aligned}A &= \textstyle\sum_{j=1}^{n} E_j, \quad E_j \in \mathcal{S} \\ B &= \textstyle\sum_{k=1}^{m} F_k, \quad F_k \in \mathcal{S}\end{aligned}$$

and $A \cap B = \emptyset$. We want to prove that $\nu(A \cup B) = \nu(A) + \nu(B)$. Trivially,

$$\begin{aligned}A \cup B &= \sum_{j=1}^{n} E_j + \sum_{k=1}^{m} F_k \\ \nu(A \cup B) &= \nu(\sum_{j=1}^{n} E_j) + \nu(\sum_{k=1}^{m} F_k) \\ &= \nu(A) + \nu(B)\end{aligned}$$

(3) (ν is unique): Recall that we have:

$$\begin{cases} \nu_1, \nu_2 : \mathcal{A}(\mathcal{S}) \to \mathbb{R}_+\{+\infty\} \\ \nu_1(A) = \nu_2(A), \quad \forall A \in \mathcal{S} \\ \nu_1, \nu_2 \text{ additive} \end{cases}$$

, and we want to prove $\nu_1(B) = \nu_2(B)$ forall $B \in \mathcal{A}(\mathcal{S})$. But this is actually very trivial, since for all $B \in \mathcal{A}(\mathcal{S})$ we have $B = \sum_{j=1}^{n} E_j$ for some $E_j \in \mathcal{S}$.

$$\nu_1(B) \overset{add}{=\!=} \sum_{j=1}^{n} \mu(E_j) \overset{add}{=\!=} \nu_2(B)$$

□

Definition 4.1.10. *infimum*

The infimum of a subset S of a partially ordered set T is the greatest element in T that is less than or equal to all elements of S, if such an element exists.

Definition 4.1.11. *outer measure*

For

$$\substack{\mathcal{C}\subseteq\mathcal{P}(\Omega)\\ \emptyset\in\mathcal{C}\\ \mu:\mathcal{C}\to\mathbb{R}_+\cup\{+\infty\}}$$

, then μ is an outer measure if

1. $\mu(\emptyset) = 0$.
2. $E \subseteq F$ where $E, F \in \mathcal{C} \Rightarrow \mu(E) \leq \mu(F)$ (monotone).
3. $E, E_i \in \mathcal{C}$, $E = \cup E_i \Rightarrow \mu(E) \leq \sum \mu(E_i)$ (sub-additive).

Now, we are going to extend the function on algebra to σ-algebra. It turns out that if a functoin $\nu : \mathcal{A} \to \mathbb{R}_+ \cup \{+\infty\}$ is σ-additive and σ-finite in algebra $\mathcal{A}$. Then we could extend it to a unique σ-additive function $\pi : \mathcal{F}(\mathcal{A}) \to \mathbb{R}_+ \cup \{+\infty\}$. ($\mathcal{F}(\mathcal{A})$ means the σ-algebra generated by $\mathcal{A}$)

There is the big picture of what we need to do to build this extension.

1. define π^* on top of an algebra $\mathcal{A}$ with a σ-additive function ν on it.

2. prove π^* is an outer meansure.

3. construct a measurealbe set $\mathcal{M}$, and prove that $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{M}$.

4. prove $\pi|_{\mathcal{M}}$ is σ-additive.

5. prove this extension is unique.

This is called caratheodory theorem.

First we are going to define the π^* on $\mathcal{P}(\Omega) \to \mathbb{R}_+ \cup \{+\infty\}$.

Definition 4.1.12. *caratheodory theorem (STEP 1)*

Suppose

$$\substack{\mathcal{A}\subseteq\mathcal{P}(\Omega),\\ \nu:\mathcal{A}\to\mathbb{R}_+\cup\{+\infty\},} \quad \substack{\text{algebra}\\ \sigma\text{-additive}}$$

, then we define π^* as

$$\pi^*(A) = \inf_{\{E_i\}} \sum \nu(E_i)$$

where $\{E_i\}$ enumerates all the countable sequence where $E_i \in \mathcal{A}$ and $A \subseteq \cup E_i$.

Lemma 4.1.3. *π^* is an outer measure*

Which means we need to prove that

1. $\pi^*(\emptyset) = 0$
2. $E \subseteq F \Rightarrow \pi^*(E) \leq \pi^*(F)$ *(monotone).*
3. $E \subseteq \cup E_i \Rightarrow \pi^*(E) \leq \sum_{i \geq 1} \pi^*(E_i)$ *(sub-additive)*

Proof. *(1) ($\pi^*(\emptyset) = 0$):*

Since

$$\underset{E_i=\emptyset}{\{E_i\}} \Rightarrow \pi^*(\emptyset) \leq \sum_{i \geq 1} \nu(E_i) = 0$$

and

$$\underset{\substack{E_i \in \mathcal{A} \\ \emptyset \subseteq \cup E_i}}{\{E_i\}} \Rightarrow \sum_{i \geq} \nu(E_i) \geq 0$$

So, we have $\pi^*(\emptyset) = 0$.

(2) ($E \subseteq F \Rightarrow \pi^*(E) \leq \pi^*(F)$):

By definition:

$$\begin{aligned} \pi^*(E) &= \inf_A \textstyle\sum_{i \geq 1} \nu(E_i), \quad & A = \{\{E_i\} | E \subseteq \cup E_i\} \\ \pi^*(F) &= \inf_B \textstyle\sum_{i \geq 1} \nu(F_i), \quad & B = \{\{F_i\} | F \subseteq \cup F_i\} \end{aligned}$$

And we could see that since $E \subseteq F$, then $B \subseteq A$. So the infimum in A should less or equal than the infimum in B, which means

$$\pi^*(E) \leq \pi^*(F)$$

(3) ($E \subseteq \cup E_i \Rightarrow \pi^*(E) \leq \sum_{i \geq 1} \pi^*(E_i)$):

First, we assume that $\pi^*(E_i) < +\infty, \forall i$, since if $\pi^*(E_i) = +\infty$ for some i, then the above inequality satisfies immediately. Lets fix some $\epsilon > 0$. And since by definition

$$\pi^*(E_i) = \inf_{\{H_k\}} \sum_{k \geq 1} \nu(H_k) < +\infty$$

There exists $\{H_{i,k}\}$ where $\underset{E_i \subseteq \cup_{k \geq 1} H_{i,k}}{H_{i,k} \subseteq \mathcal{A}}$ and

$$\pi^*(E_i) \leq \sum_{k \geq 1} \nu(H_{i,k}) \leq \pi^*(E_i) + \epsilon/2^i$$

So

$$\begin{aligned} \pi^*(E) &\leq \sum_{i,k} \nu(H_{i,k}) \\ &\leq \sum_{i \geq 1} (\pi^*(E_i) + \epsilon/2^i) \\ &= \sum_i pi^*(E_i) + \epsilon \end{aligned}$$

Q: why we need to use a ϵ statement here?

$\pi^*(E)$ is the infimum of $\{\sum_{i \geq 1} E_i | \{E_i\}\}$, thus, we may have the condition where

$$\pi^*(E) \notin \{\sum_{i \geq 1} E_i | \{E_i\}\}$$

Since this is true for all ϵ, we have $\pi^*(E) \leq \sum_{i\geq 1} \pi^*(E_i)$.

Now we define the measurealbe subset $\mathcal{M}$. □

Definition 4.1.13. *measurealbe set $\mathcal{M}$*

$A \in \mathcal{M}$ if $\forall E \subseteq \Omega, \pi^*(E) = \pi^*(E \cap A) + \pi^*(E \cap A^c)$

Observation 4.1.3.

For any $A \subseteq \Omega$ we have $pi^*(E) \leq \pi^*(E \cap A) + \pi^*(E \cap A^c)$. So, if we want to prove $\pi^*(E) = \pi^*(E \cap A) + \pi^*(E \cap A^c)$ later, we only need to shouw that $\pi^*(E) \geq \pi^*(E \cap A) + \pi^*(E \cap A^c)$.

Fact 4.1.1.

$\mathcal{A} \subseteq \mathcal{M}$.

Proof. For all the set $A \in \mathcal{A}$, we need to prove that for all $E \subseteq \Omega$, $\pi^*(E) \geq \pi^*(E \cap A) + \pi^*(E \cap A^c)$.

Lets assume $\pi^*(E) \leq +\infty$. Then, for some fixed $\epsilon > 0$, there exists $\underset{\substack{E_i \in \mathcal{A} \\ E = \cup_{i\geq 1} E_i}}{\{E_i\}}$, such that $\pi^*(E) \leq \sum_{i\geq 1} \nu(E_i) \leq \pi^*(E) + \epsilon$.

And we have

$$\begin{aligned}
E \cap A \subseteq \cup_{i\geq 1} E_i \cap A &\Rightarrow \pi^*(E \cap A) \leq \sum_{i\geq 1} \nu(E_i \cap A) \\
E \cap A^c \subseteq \cup_{i\geq 1} E_i \cap A^c &\Rightarrow \pi^*(E \cap A^c) \leq \sum_{i\geq 1} \nu(E_i \cap A^c) \\
\pi^*(E \cap A) + \pi^*(E \cap A^c) &\leq \sum_{i\geq 1} (\nu(E_i \cap A) + \nu(E_i \cap A^c)) \\
&= \sum_{i\geq 1} \nu(E_i), \quad \text{by } \sigma\text{-additive} \\
&\leq \pi^*(E) + \epsilon
\end{aligned}$$

Since this is true for all positive ϵ, we could send ϵ to 0. Then we have $\pi^*(E \cap A) + \pi^*(E \cap A^c) \leq \pi^*(E)$. □

Actually, we could not send ϵ to 0 directly. Here, what we actually do is to calculate a infimum on all possible $\{E_i\}$, and thus ϵ could reach 0.

Fact 4.1.2.

$\mathcal{M}$ is a σ-algebra.

Proof. *(1) ($\Omega \in \mathcal{M}$):*

$$\begin{aligned}
\pi^*(E \cap \Omega) + \pi^*(E \cap \Omega^c) &= \pi^*(E) + \pi^*(\emptyset) \\
&= \pi^*(E)
\end{aligned}$$

(2) ($A \in \mathcal{M} \Rightarrow A^c \in \mathcal{M}$):

This is clear, because the definition of $\mathcal{M}$ is symmetric with complement.

(3) ($\substack{\{A_j\}\\ A_j\in\mathcal{M}} \Rightarrow \cup_{j\geq 1} A_j \in \mathcal{M}$):

(3.1) ($\mathcal{M}$ is closed under finite union):

For all $A, B \in \mathcal{M}$, we want to show that $A \cup B \in \mathcal{M}$. First note that

$$\begin{aligned}
\pi^*(E \setminus A) &= \pi^*((E\setminus) \cap B) + \pi^*((E \setminus A) \setminus B) \\
&= \pi^*((E \setminus A) \cap B) + \pi^*(E \setminus (A \cup B)) \\
\pi^*(E) &= \pi^*(E \cap A) + \pi^*(E \setminus A) \\
&= \pi^*(E \cap A) + \pi^*((E \setminus A) \cap B) + \pi^*(E \setminus (A \cap B))
\end{aligned}$$

Since $(E \cap A) \cup ((E \setminus A) \cap B) = E \cap (A \cup B)$, we have, by definition

$$\pi^*(E \cap (A \cup B) \leq \pi^*(E \cap A) + \pi^*((E \setminus A) \cap B)$$

So, we have

$$\pi^*(E) \geq \pi^*(E \cap (A \cup B)) + \pi^*(E \setminus (A \cup B))$$

(3.2) ($\mathcal{M}$ is closed under countable union):

Suppose we have $\substack{\{A_j\}\\ A_j\in\mathcal{M}\\ A=\cup_{j\geq 1}A_j}$, we want to show that, $\forall E \in \mathcal{M}, \pi^*(E) \geq \pi^*(E \cap A) + \pi^*(E \cap A^c)$. Since we have proved that $\mathcal{M}$ is closed under finite union, for some fixed n, we have

$$\begin{aligned}
\pi^*(E) &= \pi^*(E \cap \bigcup_{j=1}^{n} A_j) + \pi^*(E \setminus (\bigcup_{j=1}^{n} A_j)) \\
&\geq \pi^*(E \cap \bigcup_{j=1}^{n} A_j) + \pi^*(E \setminus A), \quad \text{by monotone}
\end{aligned}$$

Now, lets define

$$\begin{aligned}
F_1 &= A_1 \\
F_2 &= A_2 \setminus A_1 \\
&\vdots \\
F_n &= A_n \setminus (A_1 \cup A_2 \cup \cdots \cup A_{n-1})
\end{aligned}$$

Then we have $\cup_{j=1}^{n} A_j = \cup_{j=1}^{n} F_j$ and $F_i \cap F_j = \emptyset$. So, now we have

$$\pi^*(E) \geq \pi^*(E \cap \sum_{j=1}^{n} F_j) + \pi^*(E \setminus A)$$

Claim 4.1.3. $\pi^*(E \cap \sum_{j=1}^{n} F_j) = \sum_{j=1}^{n} \pi^*(E \cap F_j)$

Proof. We prove this claim by induction. It is clear that when $n = 1$, we have noting to prove. Then, assume that

$$\pi^*(E \cap \sum_{j=1}^{n} F_j) = \sum_{j=1}^{n} \pi^*(E \cap F_j)$$

for some n. we want to show

$$\pi^*(E \cap \sum_{j=1}^{n+1} F_j) = \sum_{j=1}^{n+1} \pi^*(E \cap F_j)$$

. Note that

$$\begin{aligned} \pi^*(E \cap \sum_{j=1}^{n+1} F_j) &= \pi^*(E \cap \sum_{j=1}^{n+1} F_j \cap F_{n+1}) + \pi^*(E \cap \sum_{j=1}^{n+1} F_j \cap F_{n+1}^c) \\ &= \pi^*(E \cap F_{n+1}) + \pi^*(E \cap \sum_{j=1}^{n} F_j) \end{aligned}$$

The last step holds because F_j are disjoint. □

Using this claim, we have

$$\begin{aligned} \pi^*(E) &\geq \sum_{j=1}^{n} \pi^*(E \cap F_j) + \pi^*(E \setminus A) \\ \pi^*(E) &\geq \sum_{j \geq 1} \pi^*(E \cap F_j) + \pi^*(E \setminus A) \\ &\geq \pi^*(E \cap A) + \pi^*(E \setminus A) \end{aligned}$$ □

Remark 4.1.2. *Since $\mathcal{M}$ is a σ-algebra, and it contains algebra $\mathcal{A}$, we have $\mathcal{F}(\mathcal{A}) \subseteq \mathcal{M}$.*

Lemma 4.1.4.

$$\pi^*|_{\mathcal{M}} : \mathcal{M} \to \mathbb{R}_+ \cup \{+\infty\}, \qquad \textit{is } \sigma\textit{-additive}$$
$$\pi^*(A) = \nu(A), \qquad \forall A \in \mathcal{A}$$

Proof. *(1) ($\pi^*(A) = \nu(A), \forall A \in \mathcal{A}$):*

(1.1) ($\pi^*(A) \leq \nu(A)$):

Let $E_1 = A, E_2 = \emptyset, E_3 = \emptyset, \cdots$, then we have

$$\pi^*(A) \leq \sum_{j \geq 1} E_j = \nu(A)$$

(1.2) ($\nu(A) \leq \sum_{j \geq 1} \nu(E_j)$ for all $\substack{E_j \in \mathcal{A} \\ A \subseteq \cup_{j \geq 1} E_j}$):

Now, lets define

$$\begin{aligned} F_1 &= E_1 \\ F_2 &= E_2 \setminus E_1 \\ &\vdots \\ F_n &= E_n \setminus (E_1 \cup E_2 \cup \cdots \cup E_{n-1}) \end{aligned}$$

Then we have $\cup_{j=1}^{n} E_j = \cup_{j=1}^{n} F_j$ and $F_i \cap F_j = \emptyset$. So

$$\begin{aligned} A &\subseteq \cup_{j\geq 1} F_j \\ A &= \sum_{j\geq 1} F_j \cap A \\ \nu(A) &= \sum_{j\geq 1} \nu(F_j \cap A), \quad \nu \text{ is } \sigma\text{-additive on } \mathcal{A} \\ &\leq \sum_{j\geq 1} \nu(E_j), \quad \text{since } F_j \cap A \subseteq E_j \end{aligned}$$

Since we consider this for all $\{E_j\}$, we have:

$$\begin{aligned} \nu(A) &\leq \inf_{\{E_j\}} \sum_{j\geq 1} \nu(E_j), \quad \{\{E_j\}|_{E_j\in\mathcal{A}, A\subseteq \cup_{j\geq 1} E_j}\} \\ &= \pi^*(A) \end{aligned}$$

(2) ($\pi^*|_{\mathcal{M}}$ is σ-additive):

Recall that we want to prove

$$\substack{A_j\in\mathcal{M} \\ A_j\cap A_k=\emptyset} \Rightarrow \pi^*(\sum_{j\geq 1} A_j) = \sum_{j\geq 1} \pi^*(A_j)$$

First, by sub-additive of π^*, we know that

$$\pi^*(\sum_{j\geq 1} A_j) \leq \sum_{j\geq 1} \pi^*(A_j)$$

On the other hand, we have

$$\begin{aligned} \pi^*(\sum_{j\geq 1} A_j) &\geq \pi^*(\sum_{j=1}^{n} A_j) \quad \text{monotone} \\ &= \sum_{j=1}^{n} \pi^*(A_j) \quad \text{additive proved by Claim 4.1.3} \\ \pi^*(\sum_{j\geq 1} A_j) &\geq \sum_{j\geq 1} \pi^*(A_j) \end{aligned}$$

So, we have

$$\pi^*(\sum_{j\geq 1} A_j) = \sum_{j\geq 1} \pi^*(A_j)$$

□

Now we want to show that π^* is unique, according to ν. At first, we need to introduce some tools.

Definition 4.1.14. *monotone class*

For $\mathcal{G} \subseteq \mathcal{P}(\Omega)$, $\mathcal{G}$ is a monotone class if

1. $\substack{\{A_j\}\\ A_j\in\mathcal{G}\\ A_j\subseteq A_{j+1}} \Rightarrow A = \cup_{j\geq 1} A_j \in \mathcal{G}$
2. $\substack{\{B_j\}\\ B_j\in\mathcal{G}\\ B_j\supseteq B_{j+1}} \Rightarrow B = \cap_{j\geq 1} B_j \in \mathcal{G}$

Remark 4.1.3. *Since it is easy to verify that the intersection of two monotone class is also a monotone class, we could define the monotone class generated by class $\mathcal{C}$ by $\mathcal{M}(\mathcal{C})$.*

Now we give a lemma that will be proved in the latter sections.

Lemma 4.1.5.

If $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is an algebra, then $\mathcal{M}(\mathcal{A}) = \mathcal{F}(\mathcal{A})$.

Definition 4.1.15. *σ-finite*

If we say Ω is σ-finite on μ, then we means:
$\substack{\{E_j\}\\ E_j\subseteq\Omega\\ E_j\uparrow\Omega} \Rightarrow \mu(E_j) < +\infty, \forall j.$

Remark 4.1.4. *Under this setting, $\mu(\Omega)$ could be $+\infty$, but any sequence converges to Ω should be finite.*

Lemma 4.1.6. *uniqueness*

We require that

$$\substack{\mu_1,\mu_2:\mathcal{F}(\mathcal{A})\to\mathbb{R}_+\cup\{+\infty\},\\ \mu_1|_{\mathcal{A}}=\mu_2|_{\mathcal{A}}} \qquad \sigma\text{-additive}$$

and μ_1 is σ-finite by only consider the sequences $\substack{\{E_j\}\\ E_j\in\mathcal{A}\\ E_j\uparrow\Omega}$ (This also means that μ_2 is σ-finite because μ_1 and μ_2 are coincident in $\mathcal{A}$). Then $\mu_1 = \mu_2$.

Proof. Take a sequence which satisfies $\substack{\{E_j\}\\ E_j\in\mathcal{A}\\ \Omega=\cup_{j\geq 1}E_j\\ \mu_1(E_j)<+\infty}$, We could define a sequence of sets on top of it.

Actually, by σ-finite, all the sequence $\{E_j\}$ such that $E_j \uparrow \Omega$ have the property that $\mu_1(E_j) \leq +\infty$. Here, the $\Omega \in \mathcal{F}(\mathcal{A})$, so that we could add the restriction $E_j \in \mathcal{A}$ and still have $\Omega = \cup_{j\geq 1} E_j$.

$$\mathcal{B}_n = \{E \in \mathcal{F}(\mathcal{A}) | \mu_1(E \cap E_n) = \mu_2(E \cap E_n)\}$$

.

(1) ($\mathcal{B}_n \supseteq \mathcal{A}$):

For $E \in \mathcal{A}$, we have $E \cap E_n \subseteq A$, so, by definition

$$\mu_1(E \cap E_n) = \mu_2(E \cap E_n)$$

.

(2) ($\mathcal{B}_n$ is a monotone class):

(2.1) ($\substack{\{A_j\} \\ A_j \in \mathcal{B}_n \\ A_j \subseteq A_{j+1} \\ A = \cup_{j \geq 1} A_j}$ $\Rightarrow A \in \mathcal{B}_n$):

By definition, $\mu_1(A_j \cap E_n) = \mu_2(A_j \cap E_n)$. And, since μ_1, μ_2 are σ-additive, they are continue from below, which means

$$\begin{aligned} A_j \cap E_n \uparrow A \cap E_n &\Rightarrow \mu_1(A_j \cap E_n) \uparrow \mu_1(A \cap E_n) \\ &\Rightarrow \mu_2(A_j \cap E_n) \uparrow \mu_2(A \cap E_n) \end{aligned}$$

And since in each j, $\mu_1(A_j \cap E_n) = \mu_2(A_j \cap E_n)$, so their limit are also equal. Thus

$$\mu_1(A \cap E_n) = \mu_2(A \cap E_n)$$

So, $A \in \mathcal{B}_n$.

(2.2) ($\substack{\{B_j\} \\ B_j \in \mathcal{B}_n \\ B_j \supseteq B_{j+1} \\ B = \cap_{j \geq 1} B_j}$ $\Rightarrow B \in \mathcal{B}_n$):

By definition, $\mu_1(B_j \cap E_n) = \mu_2(B_j \cap E_n)$. Since $\mu_1(E_n)$ is finite, so we have

$$\begin{aligned} B_j \cap E_n \downarrow B \cap E_n &\Rightarrow \mu_1(B_j \cap E_n) \downarrow \mu_1(B \cap E_n) \\ &\Rightarrow \mu_2(B_j \cap E_n) \downarrow \mu_2(B \cap E_n) \end{aligned}$$

Bnd since in each j, $\mu_1(B_j \cap E_n) = \mu_2(B_j \cap E_n)$, so their limit are also equal. Thus

$$\mu_1(B \cap E_n) = \mu_2(B \cap E_n)$$

So, $B \in \mathcal{B}_n$.

Put them together, we have $\mathcal{B}_n$ is a monotone class. So we have $\mathcal{B}_n \supseteq \mathcal{M}(\mathcal{A}) = \mathcal{F}(\mathcal{A})$. But by definition, $\mathcal{B}_n$ is also contained in $\mathcal{F}(\mathcal{A})$, so $\mathcal{B}_n = \mathcal{F}(\mathcal{A})$.

(3) ($\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{F}(\mathcal{A})$):

Since $A \in \mathcal{F}(\mathcal{A})$, we have $A \in \mathcal{B}_n$. So $\mu_1(A \cap E_n) = \mu_2(A \cap E_n)$. Since μ_1 and μ_2 are σ-additive, they continuous from below, So we have $\mu_1(A \cap \Omega) = \mu_2(A \cap \Omega)$, so $\mu_1(A) = \mu_2(A)$. □

5 Index