Stochastic Processes

Muchang Bahng

Spring 2023

Contents

1 Introduction **3**
1.1 Transitioning from Discrete to Continuous State Space 4

2 Discrete-Time Markov Processes **6**
2.1 Classification of States 11
2.1.1 Stopping Time and Strong Markov Property 11
2.1.2 Irreducibility 12
2.1.3 Periodicity 14
2.2 Stationary Measures 14
2.2.1 Uniqueness 16
2.2.2 Reversed Markov Process 16
2.3 Reversibility (Detailed Balance) 18
2.3.1 Metropolis-Hastings Algorithm 19
2.3.2 Kolmogorov Cycle Condition 20
2.4 Ergodicity 20

3 Poisson Processes **22**
3.1 Exponential Distribution 22
3.2 Defining the Poisson Process 23
3.3 Constructing the Poisson Process 24

4 Continuous-Time Markov Processes **24**
4.1 Generator 28
4.2 Classification of States 30
4.2.1 Holding Times and Jumping Times 30
4.2.2 Irreducibility 31
4.3 Stationary Measures 31
4.3.1 Uniqueness 33
4.3.2 Reversed Markov Process 33
4.4 Reversibility (Detailed Balance) 35
4.5 Ergodicity 36

5 Martingales **37**

6 Concentration Inequalities **39**
6.1 Talagrand's Gaussian Inequality 41

7 Variance Bounds and Poincare Inequalities **43**
7.1 Markov Semigroups 52
7.2 Poincare Inequalities 54
7.2.1 The Gaussian Poincare Inequality 55

7.3 Variance Identities and Exponential Ergodicity 59

8 Subgaussian Concentration and log-Sobolev Inequalities 60
8.1 Subgaussian Variables and Chernoff Bounds 60
8.2 The Martingale Method 65
8.3 The Entropy Method 68
8.4 Modified log-Sobolev Inequalities 70

9 Lipschitz Concentration and Transportation Inequalities 71
9.1 Concentration in Metric Spaces 71

1 Introduction

Ordinary differential equations model deterministic systems that can be solved exactly through integration. For example, consider the population model determined by a linear DEQ

$$\frac{dN}{dt} = \alpha(t)N(t)$$

where N is the population size and α is a growth rate. Then, we can solve with analysis by integrating the following with a change of basis

$$\int \frac{1}{N(t)} \frac{dN}{dt}\, dt = \int \alpha(t)\, dt \iff \int \frac{1}{N}\, dN = \int \alpha(t)\, dt$$
$$\iff N(t) = C \exp\left(\int \alpha(t)\, dt \right)$$

This classical exponential growth model is not only continuous, but *smooth*, and it is this smoothness that allows us to do calculus on it. But more realistic models will have noise, which can be modeled by a random variable. Let $\alpha = r + \eta$, where r is the deterministic term and η is the random term. Then, integrating gives us

$$\frac{dN}{dt} = \big(r(t) + \eta(t)\big)N(t) \iff \int \frac{1}{N} \frac{dN}{dt}\, dt = \int r(t)\, dt + \int \eta(t)\, dt$$

The first integral can be evaluated, but classical calculus does not allow us to integrate the random part. This is where stochastic calculus is needed. Now recall from probability that a random variable over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is simply a $\mathcal{F}$-measurable function X. As some warm up exercises, let us prove a few examples.

Example 1.1 (Class 1)

Example 1.2 (Class 2)

Definition 1.1 (Stochastic Process)

A **stochastic process** is a collection of random variables indexed by time $\{X_t\}_{t \in T}$ with their respective measures ρ_t.

1. If T is countable (usually integers), then it is called a **discrete-time** stochastic process.
2. If T is continuous, then it is called a **continuous-time** stochastic process.

It is also good to think of it as a probability distribution over a space of paths.

We first start off with Markov processes. We can divide them into four kinds, depending on whether we are using discrete or continuous time, and whether we are using discrete or continuous state space. Since process over continuous state space is a natural generalization of those in a discrete one, we only distinguish between the times. When talking about continuous time, there are additional operators we must introduce, such as generators. Before we go any further, I would like to mention that these set of notes will write down the transition matrices of Markov chains as left-stochastic matrices, as they are usually written in convention. Therefore, a transition matrix would look like

$$\mathbb{P} = \begin{pmatrix} P(1,1) & \dots & P(d,1) \\ \vdots & \ddots & \vdots \\ P(1,d) & \dots & P(d,d) \end{pmatrix}$$

where $P(i,j)$ represents the probability of transition from state i to state j. Therefore, the rows must sum to 1. I use this notation because it is consistent with when we are working with Markov processes over general measurable state spaces. Note that we will denote in math font general objects and operators (X_t, ρ_t, P_s, π) and their realization as vectors and matrices in bold font $(\boldsymbol{\rho_t}, \mathbf{P_s}, \boldsymbol{\pi})$.

1.1 Transitioning from Discrete to Continuous State Space

Let us remind ourselves of the definitions involving Markov chains over a discrete state space. Let X_t be the state at time t. The discrete distribution of X_t can be represented as a column vector $\boldsymbol{\rho_t}$, where $\boldsymbol{\rho_t}(i) = \mathbb{P}(X_t = i)$, and we can calculate the distribution of X_{t+s} as

$$\boldsymbol{\rho}_{\boldsymbol{t+s}}^T = \boldsymbol{\rho}_{\boldsymbol{t}}^T \boldsymbol{P_s}$$

where $\boldsymbol{P_s}$ is a stochastic matrix. Note that representing a discrete measure on discrete $S = \{1, \dots, d\}$ with a vector really just a notational convenience for computations. We must properly distinguish the three:

1. the actual state X_t
2. the probability distribution ρ_t, which is a measure
3. the PMF vector $\boldsymbol{\rho_t}$, which is just a convenient representation of ρ_t in the way that

$$\boldsymbol{\rho_t}(i) = \rho_t(\{i\}) = \mathbb{P}(X_t = i)$$

That is, the ith element is just the measure on the singleton set $\{i\} \in \mathcal{S} = 2^S$.

The PMF vector $\boldsymbol{\rho_t}$ is really just a way to describe X_t and its distribution, which is redundant. Furthermore, when we try to describe states X_t in general measure spaces $(S, \mathcal{S})$, we cannot think of it as a vector anymore. This is not a problem in even countable spaces since we can just assign $\boldsymbol{\rho_t}(i) = \mathbb{P}(X_t = i)$ in a finite space, but for uncountably infinite spaces we cannot do this. Therefore, we must have some measurable **function** $f : S \to \mathbb{R}$ that extracts this kind information from X_t. Therefore, we must really work with the following:

1. the actual state $X_t : (\Omega, \mathcal{F}, \mathbb{P}) \longrightarrow (S, \mathcal{S})$
2. the probability distribution ρ_t of the state X_t
3. a collection of $\mathcal{S}$-measurable functions $f : S \longrightarrow \mathbb{R}$ that describes the state

At this point, we are not sure what f is since it seems quite arbitrary. But if we fix some $A \in \mathcal{S}$ and take $f = 1_A$, then $1_A(X_t)$ encodes the information of whether X_t is in A or not. This is quite nice, since now we can think of the PMF vector $\boldsymbol{\rho_t}$ as having components defined by the functions

$$\boldsymbol{\rho_t}(i) = 1_{\{i\}}(X_t) = \mathbb{P}(X_t = i)$$

The following theorem formalizes this concept.

Theorem 1.1 ()

Two random variables $X, Y : (\Omega, \mathcal{F}, \mathbb{P}) \to (S, \mathcal{S})$ have the same distribution if

$$\mathbb{E}[f(X)] = \mathbb{E}[f(Y)]$$

for all $\mathcal{F}$-measurable $f : S \to \mathbb{R}$, which can be seen by setting $f = 1_A$ for any $A \in \mathcal{F}$.

$$\begin{aligned}\mathbb{E}[1_A(X)] = \mathbb{E}[1_A(Y)] &\implies \mathbb{P}(X \in A) = \mathbb{P}(Y \in A) \\ &\implies \mathbb{P}_X(A) = \mathbb{P}_Y(A)\end{aligned}$$

and so the measure that X and Y pushes forward to $(S, \mathcal{S})$ is precisely the same. This does not mean that they are the same random variable.

Let's talk more about f in the discrete case setting. We know that the discrete distributions are represented by a column vector. It is true that every measurable function can be written as a linear combination of simple (indicator) functions, and so in a discrete space $S = \{1, \dots, d\}$, we can write every f as

$$f = \sum_{i \in S} f_i 1_{\{i\}}$$

which outputs f_i if its input is i. We can interpret it as a column vector $\mathbf{f} = (f_1, \dots, f_d)^T$. We can see that

$$\boldsymbol{\rho}_{\boldsymbol{t}}^T \mathbf{f} = \begin{pmatrix} \boldsymbol{\rho}_{\boldsymbol{t}}(1) & \dots & \boldsymbol{\rho}_{\boldsymbol{t}}(d) \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix} = \mathbb{E}[f(X_t)]$$

and if $\mathbf{f}$ is any standard unit vector, say $(1, 0, 0)$ with $d = 3$, then

$$\boldsymbol{\rho}_{\boldsymbol{t}}^T \mathbf{f} = \begin{pmatrix} \boldsymbol{\rho}_{\boldsymbol{t}}(1) & \boldsymbol{\rho}_{\boldsymbol{t}}(2) & \boldsymbol{\rho}_{\boldsymbol{t}}(3) \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} = \mathbb{E}[1_{\{1\}}(X_t)] = \mathbb{P}(X_t = 1)$$

Therefore, every time we compute $\mathbb{E}[f(X_t)]$, we can think of it in the discrete case as dotting $\boldsymbol{\rho}_{\boldsymbol{t}}$ with a function vector $\mathbf{f}$ to extract whatever we want from the vector X_t. And as we will find out later, the linearity of the stochastic matrix $\mathbf{P_s}$ is analogous to the linearity of the Markov semigroup P_s.

Therefore, our Markov process is really just some stochastic process $\{X_t\}_{t \geq 0}$ over some measurable space $(S, \mathcal{S})$ with the property that

$$\mathbb{P}(X_{t+s} \in A \mid \{X_r \in B_r\}_{r \leq t}) = \mathbb{P}(X_{t+s} \in A \mid X_t \in B_t)$$

where $A \in \mathcal{S}$, and this captures the discrete case by setting $A = \{j\} \in 2^S$ which gives

$$\mathbb{P}(X_{t+s} = j \mid \{X_r = i_r\}_{r \leq t}) = \mathbb{P}(X_{t+s} = j \mid X_t = i_t)$$

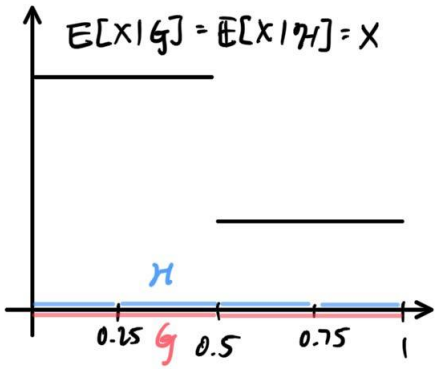
This basically says that the probability that X_{t+s} lying in A is only dependent on its present state $X_t \in B_t$, not the history $\{X_r \in B_r\}_{r \leq t}$. In fact, by using the identity $\mathbb{E}[1_A] = \mathbb{P}(A)$ and setting $f = 1_A$, we can capture this effect for *all* measurable $f : (S, \mathcal{S}) \to (\mathbb{R}, \mathcal{R})$. Thus, the Markov property now looks like

$$\mathbb{E}[f(X_{t+s}) \mid \{X_r \in B_r\}_{r \leq t}] = \mathbb{E}[f(X_{t+s}) \mid X_t \in B_t]$$

We don't need to fix the X_r's into sets B_r's and so we can write

$$\mathbb{E}[f(X_{t+s}) \mid \{X_r\}_{r \leq t}] = \mathbb{E}[f(X_{t+s}) \mid X_t]$$

Now let's talk about this Markov property. It is true that σ-algebra $\sigma(\{X_r\}_{r\leq t})$ is bigger than $\sigma(X_t)$; the Markov property does not imply that they are the same size. Rather, we should interpret this as the extra information introduced by the bigger $\sigma(\{X_r\}_{r\leq})$ is irrelevant. This is analogous to trying to approximate a function with a pointlessly large σ-algebra. For example, given a piecewise function X defined on the unit interval $\Omega = [0, 1]$, let $\mathcal{G}$ be the σ-algebra generated by $[0, 0.5), [0.5, 1]$ and $\mathcal{H}$ be that generated by $[0, 0.25), [0.25, 0.5), [0.5, 0.75), [0.75, 1]$.



Then, we can see that

$$\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{H}]$$

That is, the two random variables are exactly equal, even though $\mathcal{H}$ has more information than $\mathcal{G}$. Note that this is not the law of iterated expectations. This rule does not say that $\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}]] = \mathbb{E}[\mathbb{E}[X \mid \mathcal{H}]]$; this law is true regardless. Rather, this property is a special property of the function X, and therefore the Markov property is a special property of the stochastic process $\{X_t\}_{t\geq 0}$.

2 Discrete-Time Markov Processes

Definition 2.1 (DTMP)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ a measurable space. Then, a homogeneous **discrete-time Markov process** is a stochastic process $\{X_n\}_{n\in\mathbb{N}}$ which takes values in S (i.e. $X_n : \Omega \to S$) satisfying the **Markov property**: for every bounded measurable f and $n \geq 1$,

$$\mathbb{E}[f(X_{n+m}) \mid \{X_r\}_{r=0}^n] = \mathbb{E}[f(X_{n+m}) \mid X_n] = (P_m f)(X_n)$$

Since this is true for all n, this process is **time-homogeneous**. Note that both sides are random variables, and it says that the best estimate of $f(X_{n+m})$ as a function of $\{X_r\}_{r=0}^n$ can be simply expressed as as a function of the current X_n. Notice also that we have given a specific label $P_m f$ to the conditional expectation on the right hand side.

Since every X_n has distribution ρ_n, we can describe the entire distribution of X_n by "extracting" our desired information f with

$$\mathbb{E}[f(X_n)] = \int_S f\, \rho_n$$

Now, if we wanted to extract information f from X_{n+m}, we may not know its distribution ρ_{n+m}, but the Markov property allows us to condition X_n (which we know the distribution of) by integrating over the measure ρ_n, which we do know:

$$\mathbb{E}[f(X_{n+m}] = \mathbb{E}[\mathbb{E}[f(X_{n+m}) \mid X_n]] = \mathbb{E}[(P_m f)(X_n)] = \int_S P_m f \, \rho_n$$

So, P_m is an operator that allows us to compute anything about the distribution of X_{n+m} from the measure of X_n. That is, $\rho_{n+m}(f) = \rho_n(P_m f)$.

$$\mathbb{E}[f(X_{n+m})] = \int_S f \, \rho_{n+m} = \int_S P_m f \, \rho_n = \mathbb{E}[(P_m f)(X_n)]$$

for all measurable f. Let us now show how $P_1 = P$ realizes as a matrix in the discrete state space case.

Example 2.1 (Transition Operator as a Matrix in Discrete Space)

Given $S = \{1, \dots, d\}$, let us construct a column vector $\boldsymbol{\rho_n}$ representing the distribution of X_n. Then,

$$\begin{aligned}
\boldsymbol{\rho_{n+1}}(j) &= \mathbb{P}(X_{n+1} = j) \\
&= \mathbb{E}[1_{\{j\}}(X_{n+1})] \\
&= \mathbb{E}[\mathbb{E}[1_{\{j\}}(X_{n+1}) \mid X_n]] && = \mathbb{E}[(P1_{\{j\}})(X_n)] \\
&= \int_S \mathbb{E}[1_{\{j\}}(X_{n+1}) \mid X_n] \, d\rho_n && = \int_S P1_{\{j\}}(X_n) \, d\rho_n \\
&= \sum_{i \in S} \mathbb{P}[X_{n+1} = j \mid X_n = i] \, \mathbb{P}(X_n = i) && = \sum_{i \in S} P1_{\{j\}}(i) \, \mathbb{P}(X_n = i)
\end{aligned}$$

which can be summarized as

$$\boldsymbol{\rho_{n+1}}(j) = \sum_{i=1}^{d} P1_{\{j\}}(i) \boldsymbol{\rho_n}(i) = \sum_{i=1}^{d} \mathbb{P}(X_{n+1} = j \mid X_n = i) \, \boldsymbol{\rho_n}(i)$$

We can compactly organize the probabilities of these internode travel inside a $d \times d$ right stochastic **transition matrix**

$$\mathbf{P_t} = \begin{pmatrix} P1_{\{1\}}(1) & \dots & P1_{\{1\}}(d) \\ \vdots & \ddots & \vdots \\ P1_{\{d\}}(1) & \dots & P1_{\{d\}}(d) \end{pmatrix} = \begin{pmatrix} \mathbb{P}(X_{n+1} = 1 \mid X_n = 1) & \dots & \mathbb{P}(X_{n+1} = d \mid X_n = 1) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(X_{n+1} = 1 \mid X_n = d) & \dots & \mathbb{P}(X_{n+1} = d \mid X_n = d) \end{pmatrix}$$

and compactly write the above equation as

$$\boldsymbol{\rho}_{\boldsymbol{n+1}}^T = \boldsymbol{\rho}_{\boldsymbol{n}}^T \mathbf{P_t}$$

It immediately follows from computation that P_m is realized as $\mathbf{P}^m$, the mth power of matrix $\mathbf{P}$, which can also be shown by the Chapman-Kolmogorov equation below.

Therefore, this linear operator P_m can be seen as analogous to the probability transition matrix $\mathbf{P_m}$ of a Markov chain. We know that since they are matrices, from first glance we would guess that P_m is linear. This is indeed trivial by linearity of conditional expectation.

Lemma 2.1 ()

P_m is a linear operator. That is, for $\alpha, \beta \in \mathbb{R}$, and bounded measurable functions f, g,

$$P_m(\alpha f + \beta g) = \alpha P_m f + \beta P_m g$$

Proof.

By linearity of conditional expectation,

$$\begin{aligned}(P_m(\alpha f+\beta g))(X_n) &= \mathbb{E}[(\alpha f+\beta g)(X_{n+m}) \mid X_n] \\ &= \mathbb{E}[(\alpha f)(X_{n+m}) \mid X_n] + \mathbb{E}[(\beta g)(X_{n+m}) \mid X_n] \\ &= \alpha(Pf)(X_n) + \beta(Pg)(X_n)\end{aligned}$$

We can now interpret linearity and the Markov property in the discrete space.

Example 2.2 (Markov Property in Discrete Space)

If we wanted to extract information from X_n with function f (i.e. compute $\mathbb{E}[f(X_n)]$), we can calculate

$$\mathbb{E}[f(X_n)] = \boldsymbol{\rho}_{\boldsymbol{n}}^T \mathbf{f} = \begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

Now, say that m units of time later, we want to extract information f from X_{n+m} by computing

$$\mathbb{E}[f(X_{n+m})] = \boldsymbol{\rho}_{\boldsymbol{n+m}}^T \mathbf{f} = \begin{pmatrix}\boldsymbol{\rho_{n+m}}(1) & \dots & \boldsymbol{\rho_{n+m}}(d)\end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

The problem is that we don't know what the distribution of X_{n+m} is (i.e. don't know $\boldsymbol{\rho_{n+m}}(i)$), so we get its expectation by conditioning it on X_n, which realizes as taking the expectation of a *different* function $P_m f$ with respect to ρ_n.

$$\mathbb{E}[f(X_{n+m})] = \mathbb{E}[\mathbb{E}[f(X_{n+m}) \mid X_n]] = \mathbb{E}[(P_m f)(X_n)] = \begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \begin{pmatrix} (P_m f)_1 \\ \vdots \\ (P_m f)_d \end{pmatrix}$$

It turns out that this transformation $\mathbf{f} \mapsto \mathbf{P_m f}$ (from row vector to row vector) is linear, and so we can interpret $\mathbf{P_m}$ as $\mathbf{f}$ that has been left-multiplied by some transformation matrix $\mathbf{P_m}$.

$$\begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \begin{pmatrix} (P_m f)_1 \\ \vdots \\ (P_m f)_d \end{pmatrix} = \begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \underbrace{\begin{pmatrix} & \mathbf{P_m} & \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}}_{\mathbf{P_m f}}$$

It turns out that this $\mathbf{P_m}$ acts linearly on $\mathbf{f}$ through left multiplication, but we can also right-multiply $\boldsymbol{\rho_n}$ by $\mathbf{P_m}$ to get the new distribution of X_{n+m}!

$$\begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \begin{pmatrix} (P_m f)_1 \\ \vdots \\ (P_m f)_d \end{pmatrix} = \underbrace{\begin{pmatrix}\boldsymbol{\rho_n}(1) & \dots & \boldsymbol{\rho_n}(d)\end{pmatrix} \begin{pmatrix} & \mathbf{P_m} & \end{pmatrix}}_{\boldsymbol{\rho}_{\boldsymbol{n}}^T \mathbf{P_m} = \boldsymbol{\rho}_{\boldsymbol{n+m}}^T} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

Therefore, it turns out that the linearity of $\mathbf{P_m}$ on $\mathbf{f}$ implies linearity of it on the vector $\boldsymbol{\rho_n}$.

Now focusing on $f = 1_A$, we can define the following.

Definition 2.2 (Transition Probability)

Let us have Markov process (X_n) with operator P_m. The function $p_m : S \times \mathcal{S} \to \mathbb{R}$ defined

$$p_m(x, A) \coloneqq P_m 1_A(x) = \mathbb{E}[1_A(X_{n+m}) \mid X_n = x] = \mathbb{P}(X_{n+m} \in A \mid X_n = x)$$

is the **transition probability**, or **transition kernel**, of this chain. Note that

1. For each $x \in S$, $A \mapsto p_m(x, A)$ is a probability measure on $(S, \mathcal{S})$. This means that if we are in some place x at time n, then the probability that we will land in some subset $A \in \mathcal{S}$ of S at time $n + m$ is $p_m(x, A)$.
2. For each $A \in \mathcal{S}$, $P_m 1_A = p_m(\cdot, A)$ is a measurable function.

$$p(x, A) = \int_A p(x, y)\, dy$$

Note that by the law of total probability, we must have

$$\int_S dp(x) = 1 \text{ and } \int_S dp^{(m)}(x) = 1$$

Given that we have an initial distribution $X_0 \sim \mu_0$, we can see that the distribution $X_1 \sim \mu_1$ is defined as

$$\begin{aligned} \mathbb{P}(X_1 \in A_1) &= \int_{A_0} \mathbb{P}(X_1 \in A_1 \mid X_0 = x)\, \mathbb{P}(X_0 = x)\, dx \\ &= \int_{A_0} p(x_0, A_1)\, \mu_0(dx_0) \end{aligned}$$

Note that in the matrix realization of the example above, it looks like P_m acts on the distribution ρ_n to get a new distribution ρ_{n+m}, but this is not strictly the case since P_m is an operator on f. However, for the sake of intuitiveness, we can interpret P_m in two ways:

1. It operates on the measure ρ_n by pushing it forward in time to get ρ_{n+m}. This operator is defined as

   $$\rho_n \mapsto \rho_{n+m}(\cdot) = p_m(X_n, \cdot)$$

   which corresponds to the matrix multiplication $\boldsymbol{\rho}_{\mathbf{n}}^T \mapsto \boldsymbol{\rho}_{\mathbf{n+m}}^T = \boldsymbol{\rho}_{\mathbf{n}}^T \mathbf{P_m}$
2. It operates on the function f (at X_{n+m}) by pulling it back to $P_m f$ that operates on X_n. This operation $f \mapsto P_m f$ corresponds to the matrix multiplication $\mathbf{f} \mapsto \mathbf{P_m f}$.

Either way, we can think of the order of operations as either $(\boldsymbol{\rho}_{\mathbf{n}}^T \mathbf{P_m})\mathbf{f}$ or $\boldsymbol{\rho}_{\mathbf{n}}^T(\mathbf{P_m f})$.

Just like stochastic transition matrices, we can also deduce a semigroup property of the collection $(P_m)_{m \in \mathbb{N}}$.

Lemma 2.2 (Chapman-Kolmogorov Equation)

Given the operator P, we have

$$P_{m+k} = P_m P_k$$

which indicates

$$p_{m+k}(x, A) = \int_S p_k(x, y)\, p_m(y, A)\, dy$$

Proof.

We can compute

$$\begin{aligned} P_{m+k}f(X_n) &= \mathbb{E}[f(X_{n+m+k}) \mid X_n] \\ &= \mathbb{E}[\mathbb{E}[f(X_{n+m+k}) \mid X_{n+m}, X_n] \mid X_n] \\ &= \mathbb{E}[\mathbb{E}[f(X_{n+m+k}) \mid X_{n+m}] \mid X_n] \\ &= \mathbb{E}[Pf_k(X_{n+m}) \mid X_n] \\ &= P_m P_k f(X_n) \end{aligned}$$

Example 2.3 (Chapman-Kolmogorov in Discrete Space)

By conditioning on intermediate nodes, we can compute that

$$\mathbf{P_{m+k}}(i,j) = \sum_{s \in S} \mathbf{P_m}(i,s)\,\mathbf{P_k}(s,j) \implies \mathbf{P_{m+k}} = \mathbf{P_m P_k}$$

which can be seen by setting $x = i$ and $A = \{j\} \in 2^S$ in the transition probability above.

$$\mathbf{P_{m+k}}(i,j) = p_{m+k}(i,\{j\}) = \int_S p_m(i,\{s\})\,p_k(s,\{j\})\,ds = \sum_{s\in S} p_m(i,\{s\})\,p_k(s,\{j\}) = \sum_{s=1}^{d} \mathbf{P_m}(i,s)\mathbf{P_k}(s,j)$$

and summing this for each entry gives $\mathbf{P_{m+k}} = \mathbf{P_m P_k}$. By setting $k = 1$, an immediate consequence of this is that the m step transition probability $\mathbb{P}(X_{n+m} = j \mid X_n = i)$ is simply $\mathbf{P}^m(i,j)$, the kth power of the transition matrix $\mathbf{P}$.

We give one more property.

Lemma 2.3 (Conservativeness)

$\{P_m\}$ satisfies

$$P_m 1 = 1$$

for all $m \geq 0$, where $1 = 1_S$ is the constant function of 1.

Proof.

This is trivial since it is just the law of total probability. That is, $1_S(X_n) = 1$, and

$$(P_m 1_S)(X_n) = \mathbb{E}[1_S(X_{n+m}) \mid X_n]$$

and note that $\sigma(X_n)$ is a finer σ-algebra than that generated by $1_S(X_{n+m})$, meaning that the right hand side is equal to $1_S(X_{n+m})$ itself, which equals 1.

In discrete spaces, this property realizes into the fact that the transition matrix is stochastic, since the constantly 1 function $f = \sum_{i\in S} 1_{\{i\}}$ realizes into the $(1,\dots,1)$ vector, and

$$\begin{pmatrix} & \mathbf{P_m} & \end{pmatrix} \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$$

if and only if $\mathbf{P_m}$ is stochastic. But this is quite redundant for discrete spaces since the fact that $\mathbf{P_m}$ acts

on the indicator functions as $P_s 1_{\{j\}}(i) = \mathbb{P}(X_{t+s} = j \mid X_t = i)$ already implies that it should be stochastic (by law of total probability).

We provide with a variety of examples.

Example 2.4 (Random Walks)

A *random walk* on the integers $\mathcal{S} = \mathbb{Z}$ where a point has equal probability of moving right or left can be modeled with the probability transition matrix.

$$\mathbf{P}(i,j) = \mathbb{P}(X_{n+1} = j \,|\, X_n = i) = \begin{cases} \frac{1}{2} & j = i+1 \\ \frac{1}{2} & j = i-1 \\ 0 & otherwise \end{cases}$$

This can be generalized to multiple dimensional random walks on graphs with probability function

$$\mathbf{P}(i,j) = \frac{1}{\deg(i)}$$

where $\deg(i)$ is the number of adjacent nodes to node i. In this way, the point hops randomly from node to node, and if the graph is connected, then the walker can visit any vertex in the graph.

Example 2.5 (Discrete Moran Model)

Consider a population of size N. Each individual is one of two types (say, red or blue). At each time step, the system evolves in the following way: First, one of the individuals is chosen uniformly at random to be eliminated from the population; and another individual is chosen uniformly at random to produce one offspring identical to itself. These two choices are made independently. So, if a red individual is chosen to reproduce, and a blue one is chosen for elimination, then the total number of red particles increases by one and the number of blue particles decreases by one. If a red is chosen for reproduction and a red is chosen for elimination, then there is no net change in the number of reds and blues. Let X_n be the number of red individuals at time n. The transition matrix for this chain is

$$\mathbf{P}(j,i) = \begin{cases} \frac{i}{N}\left(\frac{N-i}{N}\right) & j = i-1, i \neq 0 \\ \left(\frac{N-i}{N}\right)\frac{i}{N} & j = i+1, i \neq N \\ 1 - 2\left(\frac{N-i}{N}\right)\frac{i}{N} & j = i \\ 0 & \text{otherwise} \end{cases}$$

Note that the states $X_n = 0$ and $X_n = N$ are absorbing states, which represents a phenomenon called *fixation*.

2.1 Classification of States

2.1.1 Stopping Time and Strong Markov Property

Definition 2.3 (Stopping Time)

Given a stochastic process $\{X_n\}$, a nonnegative integer random variable T is called a stopping time if for all integers $k \geq 0$, $T \leq k$ depends only on $X_0, \ldots, X_k$.

Example 2.6 (Coin Toss)

Let $\{X_n\}$ be a stochastic process with $X_n - X_{n-1}$ be iid standard Gaussians, with $X_0 = 0$. Then,

1. Let $T = \min\{n \geq 1 \mid X_n > 10\}$ be the first time that we surpass 10. This is a stopping time since

$$\mathbb{P}(T = k) = \mathbb{P}(X_0 \leq 10, X_1 \leq 10, \dots, X_{k-1} \leq 10, X_k > 10)$$

2. Let $T = \min\{n \geq 1 \mid X_{n+1} - X_n < 0\}$ be the time of the first peak. This is not a stopping time because you can't determine whether we have peaked at time k by looking at the X_n's up to k. You need information on X_{n+1}.
3. Let $T = \min\{n \geq 1 \mid X_n - X_{n-1} < 0\}$ be the first time we have gone down from a peak. This is a stopping time since

$$\mathbb{P}(T = k) = \mathbb{P}(X_0 < X_1 < X_2 < \dots < X_{k-1} > X_k)$$

Definition 2.4 (Time of Return)

Given a stochastic process, let the stopping time

$$T_A \coloneqq \min\{n \geq 1 \mid X_n \in A\}$$

be the random variable defined as the **time of first return to** A (being there at time $t = 0$ doesn't count). Let Let $T_A^1 = T_A$ and for $k \geq 2$,

$$T_A^k \coloneqq \min\{n > T_A^{k-1} \mid X_n \in A\}$$

be the **stopping time of the kth return to** A.

Since stopping at time k depends only on the values $X_0, \dots, X_k$, and in a Markov chain the distribution of the future only depends on the past through the current state, it should not be hard to believe that the Markov property holds at stopping times.

Theorem 2.1 (Strong Markov Property)

Suppose T is a stopping time. Then, for natural $k \geq 1$,

$$\mathbb{P}(X_{T+k} = j \mid X_T = i, \dots, X_0 = i) = \mathbb{P}(X_k = j \mid X_0 = i)$$

2.1.2 Irreducibility

Definition 2.5 (Closed Set, Absorbing State)

A set $A \subset S$ is **closed** if it is impossible to get out.

$$\mathbb{P}(X_{n+1} \in A \mid X_n \in A) = 1$$

If $A = \{i\}$ is a singleton set in some discrete state space, then i is said to be an **absorbing state**.

$$\mathbb{P}(X_{n+1} \neq i \mid X_n = i) = 0$$

Definition 2.6 (Recurrence, Transience)

A state $x \in S$ is called **recurrent** if

$$\rho_{xx} = \mathbb{P}(T_x < \infty \mid X_0 \in A) = 1$$

i.e. if the chain returns to x infinitely many times. x is said to be **transient** if $\rho_{xx} < 1$, and so eventually the Markov chain does not find its way back to x ever again.

Definition 2.7 (Communication)

We say that $x \in S$ communicates with $y \in S$, denoted $x \to y$, if

$$\rho_{xy} \coloneqq \mathbb{P}(T_y < \infty \mid X_0 = y) > 0$$

That is, there is a positive probability that we will jump from x to y in a finite amount of steps. We can also see this as there existing an $m > 0$ such that $\mathbb{P}(X_m = y \mid X_0 = x)p^m(x, y) > 0$.

Lemma 2.4 ()

The following hold.
1. If $x \to y$ and $y \to z$, then $x \to z$.
2. If $\rho_{xy} > 0$ but $\rho_{yx} = 0$, then x is transient.
3. If x is recurrent and $\rho_{xy} > 0$, then $\rho_{yx} = 1$.

Definition 2.8 (Irreducible Set)

A set $B \subset S$ is called **irreducible** if for all $i, j \in B$, i communicates with j.

Theorem 2.2 ()

If C is a finite closed and irreducible set, then all states in C are recurrent.

Theorem 2.3 (Decomposition)

If the state space S is finite, then S can be written as a disjoint union

$$T \cup R_1 \cup \ldots \cup R_k$$

where T is a set of transient states and R_i are closed irreducible sets of recurrent states.

Lemma 2.5 ()

If x is recurrent and $x \to y$, then y is recurrent.

Lemma 2.6 ()

In a finite closed set there has to be at least one recurrent state.

2.1.3 Periodicity

Definition 2.9 (Period)

For any state $x \in \mathcal{S}$, the **period** of x is defined to be

$$d(x) \equiv \gcd\{n \geq 1 \mid P^{(n)}(x,x) > 0\}$$

Lemma 2.7 ()

If $p(x,x) > 0$ (not $\rho_{xx} > 0$!), then x has period 1.

Theorem 2.4 ()

If two states x and y communicate, then they must have the same period

$$d(x) = d(y)$$

It naturally follows that if $B \subset S$ is irreducible, then all states must have the same period.

Definition 2.10 ()

If an irreducible chain has period 1, the chain is said to be **aperiodic**. Otherwise, the chain is *periodic* with period $d > 1$.

2.2 Stationary Measures

Recall that a discrete time Markov process $(X_n)_{n\in\mathbb{N}}$ evolves, and this evolution can be described by the sequence of measures $(\rho_n)_{n\geq 0}$ for each X_n. If we would like to measure X_{n+m} with function f, we can calculate $\mathbb{E}[f(X_{n+m})] = \mathbb{E}_{\rho_{n+m}}[f]$, but we don't know ρ_{n+m}. Fortunately, we can "pull back" the f to compute the equivalent

$$\mathbb{E}_{\rho_{n+m}}[f] = \mathbb{E}[f(X_{n+m})] = \mathbb{E}[\mathbb{E}[f(X_{n+m}) \mid X_n]] = \mathbb{E}[P_m f(X_n)] = \mathbb{E}_{\rho_n}[P_m f]$$

which essentially measures X_{n+m} with f by measuring X_n with $P_m f$. Now, we want to construct a stationary measure μ that captures the fact that if a certain state $X_n \sim \rho_n = \mu$, then the measure of future $X_{n+m} \sim \rho_{n+m} = \mu$ also. If μ is stationary, then both $\rho_{n+m} = \rho_n = \mu$, and this is equivalent to

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_m f]$$

for all measurable f and $m \geq 0$. This will be the definition that we will work with. To help with the interpretation, we can restrict the case to $f = 1_A$ to get $\mathbb{P}(X_n \in A) = \mathbb{P}(X_{n+m} \in A)$ for all $A \in \mathcal{S}$, which means that the probability of X_{n+m} realizing in A is equal to the probability of X_n realizing in A. In summary, stationary measures describe the equilibrium or steady-state behavior of the Markov process.

Definition 2.11 (Stationary Measure)

A probability measure μ is called **stationary** or **invariant** if

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_m f], \text{ conventionally written as } \mu(f) = \mu(P_m f)$$

for all $m \geq 0$ and bounded measurable f. This is a property of the *measure*.

To give a pictorial interpretation, imagine an initial distribution $X_0 \sim \rho_0$ as some amount of sand placed on the state space S (either as a continuous mass or mounds on discrete nodes). After one step, the distribution will evolve to $X_1 \sim \rho_1$, where a different mound of sand will form on S. If $\rho_0 = \mu$, then the flow of sand between the nodes will balance each other out, and we still have the same amount of sand $\rho_1 = \mu$ after each step. The discrete case is simpler, since we can just imagine there being $\boldsymbol{\pi}(i)$ of sand at node i, and $\mathbf{P}(i,j)$ of its proportion of sand flowing from node i to j at each step. Therefore, all the sand flowing out of i, which is $\sum_{j=1}^d P(i,j)\boldsymbol{\pi}(i) = 1$, balances out with the flow of sand into i, which is $\sum_{j=1}^d P(j,i)\boldsymbol{\pi}(j)$.

$$1 = \sum_{i=1}^d P(i,j)\boldsymbol{\pi}(i) = \sum_{j=1}^d P(j,i)\boldsymbol{\pi}(j)$$

and doing this for all i realizes into the matrix equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P}$.

Example 2.7 (Stationary Distribution in Discrete Space)

Given discrete state space $S = \{1, \dots, d\}$, our stationary measure μ can be represented by the all familiar vector

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}(1) & \dots & \boldsymbol{\pi}(d) \end{pmatrix} = \begin{pmatrix} \mu(\{1\}) & \dots & \mu(\{d\}) \end{pmatrix}$$

Given the PMF vectors $\boldsymbol{\rho_n} = \boldsymbol{\pi}$ and $\boldsymbol{\rho_{n+m}} = \boldsymbol{\pi}$ and some measurable function $\mathbf{f} = (f_1, \dots, f_d)^T$, the stationary distribution property says that

$$\mathbb{E}[f(X_{n+s})] = \mathbb{E}[(P_m f)(X_n)] \iff \boldsymbol{\pi}\mathbf{f} = \boldsymbol{\pi}\mathbf{P_m}\mathbf{f}$$

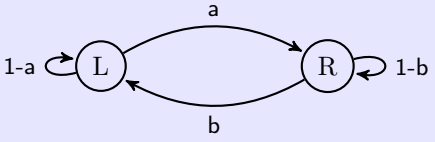
which means that $\mathbf{P_m}\mathbf{f}$ will act on $\boldsymbol{\pi}$ the same way that $\mathbf{f}$ does (though $\mathbf{P_m}\mathbf{f} \neq \mathbf{f}$). We can also interpret $\boldsymbol{\pi}$ as the eigenvector of $\mathbf{P}$ with eigenvalue 1, so that it is invariant.

Example 2.8 (Two Node System)

Let us have a two node system with nodes labeled L and R. That is, $\mathcal{S} = \{L, R\}$. Consider a chain on this state space with transition probability matrix.

$$\mathbf{P} = \begin{pmatrix} 1-a & a \\ b & 1-b \end{pmatrix}$$

which can be visualized in the following diagram below.



Then, the stationary distribution is

$$\boldsymbol{\pi} = \left(\frac{b}{a+b}, \frac{a}{a+b} \right)$$

Notice that if $a = b = 0$, then this definition is ill-defined, and any probability distribution is invariant since $P = I_2$, the identity matrix.

This is also stationary since with certain conditions, the limiting behavior of the chain converges to π, but we will prove that later.

Definition 2.12 (Doubly Stochastic Chains)

A transition matrix $\mathbf{P}$ is said to be **doubly stochastic** if its columns also sum to 1.

Theorem 2.5 ()

Given a Markov chain with state space $S = \{1, \dots, d\}$, its transition probability matrix $\mathbf{P}$ is doubly stochastic if and only if its stationary distribution is the uniform distribution

$$\boldsymbol{\pi} = \left(\frac{1}{d}, \frac{1}{d}, \dots, \frac{1}{d} \right)$$

Proof.

We prove the only if part. Let $\pi(i) = 1/N$ for all $i = 1, \dots, N$. Then, for $j = 1, \dots, N$,

$$(\boldsymbol{\pi}\mathbf{P})(i) = \sum_{j=1}^{N} \pi(j)\mathbf{P}(j,i) = \frac{1}{N} \sum_{j=1}^{N} \mathbf{P}(j,i) = \frac{1}{N} = \pi(i)$$

The if part is very similar.

2.2.1 Uniqueness

TBD TBD

2.2.2 Reversed Markov Process

From now, given the state space $(S, \mathcal{S})$ we can put a measure μ on it to get a measure space $(S, \mathcal{S}, \mu)$. The Banach space of all μ-measurable functions $f : (S, \mathcal{S}, \mu) \to (\mathbb{R}, \mathcal{R})$ (i.e. for every Borel $B \in \mathcal{R}$, $f^{-1}(B) \in \mathcal{S}$) will be denoted $L^p(\mu)$, equipped with the norm

$$||f||_{L^p(\mu)} \coloneqq \mathbb{E}_\mu[f^p]^{1/p} = \left(\int_S |f|^p \, d\mu \right)^{1/p}$$

If $p = 2$, then we can define the inner product

$$\langle f, g \rangle_\mu \coloneqq \mathbb{E}_\mu[fg] = \int_S fg \, d\mu$$

Lemma 2.8 (Contraction of Stationary Measure)

Let μ be a stationary measure. Then,

$$||P_t f||_{L^p(\mu)} \geq ||f||_{L^p(\mu)} = \mathbb{E}_\mu[f^p]^{1/p}$$

Now, we can construct reversed Markov processes.

Definition 2.13 (Reversed Markov Process)

Let $\{X_n\}_{n=0}^N$ be a discrete time Markov process with transition operator $P = P_1$ (and semigroup $(P_m = P^m)$) and stationary distribution μ. Then, fix N and let $Y_n = X_{N-n}$. Then, Y_n is a discrete time Markov process with the **dual transition operator** P^*, the adjoint of P satisfying

$$\langle f, Pg \rangle_\mu = \langle P^* f, g \rangle_\mu$$

for all bounded measurable $f, g \in L^2(\mu)$.

Though we have given the reversed Markov process as a definition above, we can prove that this satisfies the Markov property.

Proof.

We can see how this definition realizes in a discrete space.

Example 2.9 ()

Given $S = \{1, \dots, d\}$ and function vectors $\mathbf{f}, \mathbf{g}$,

$$\langle f, g \rangle_\mu = \int_S f g d\mu = \sum_{i=1}^d f_i g_i \pi(i)$$

and by definition of the adjoint, we must have

$$\begin{aligned} \langle f, Pg \rangle_\mu = \sum_{i=1}^d f_i (\mathbf{Pg})_i \pi(i) &= \sum_{i=1}^d f_i \left(\sum_{j=1}^d \mathbf{P}(i,j) g_j \right) \pi(i) \\ &= \sum_{i=1}^d g_i \left(\sum_{j=1}^d \mathbf{P}^*(i,j) f_j \right) \pi(i) = \sum_{i=1}^d (\mathbf{P}^* \mathbf{f})_i \, g_i \, \pi(i) = \langle P^* f, g \rangle_\mu \end{aligned}$$

A bit of computation will show us that

$$\mathbf{P}^*(i,j) = \frac{\mathbf{P}(j,i)\pi(j)}{\pi(i)}$$

and we can indeed check that

$$\begin{aligned} \langle P^* f, g \rangle_\mu &= \sum_{i=1}^d g_i \left(\sum_{j=1}^d \mathbf{P}^*(i,j) f_j \right) \pi(i) \\ &= \sum_{i=1}^d g_i \left(\sum_{j=1}^d f_j \frac{\mathbf{P}(j,i)\pi(j)}{\pi(i)} \right) \pi(i) \\ &= \sum_{j=1}^d \sum_{i=1}^d g_i \, f_j \mathbf{P}(j,i) \, \pi(j) \\ &= \sum_{j=1}^d f_j \left(\sum_{i=1}^d g_i \mathbf{P}(j,i) \right) \pi(j) \\ &= \sum_{j=1}^d f_j (\mathbf{Pg})_j \pi(j) = \langle f, Pg \rangle_\mu \end{aligned}$$

Note that $\mathbf{P}^*$ also satisfies $\mathbf{P}^*(i,j) \geq 0$ and by definition of the stationary distribution π,

$$\sum_{j=1}^d \mathbf{P}^*(i,j) = \sum_{j=1}^d \frac{\mathbf{P}(j,i)\pi(j)}{\pi(i)} = \frac{1}{\pi(i)} \sum_{j=1}^d \mathbf{P}(j,i)\pi(j) = \frac{\pi(i)}{\pi(i)} = 1$$

Note that the transition probability is computed using Bayes rule

$$\begin{aligned}\mathbf{P}^*(i,j) &= \mathbb{P}(Y_{m+1}=j \mid Y_m = i) \\ &= \frac{\mathbb{P}(Y_m = i \mid Y_{m+1}=j)\mathbb{P}(Y_{m+1}=j)}{\mathbb{P}(Y_m=i)} \\ &= \frac{\mathbb{P}(X_{n-m}=i \mid X_{n-m-1}=j)\mathbb{P}(X_{n-m-1}=j)}{\mathbb{P}(X_{n-m}=i)} \\ &= \frac{\mathbf{P}(j,i)\pi(j)}{\pi(i)}\end{aligned}$$

and $\{Y_m\}$ also satisfies the Markov property.

$$\begin{aligned}&\mathbb{P}(Y_{m+1}=j \mid Y_m=i, Y_{m-1}=i_{m-1},\dots,Y_0=i_0) \\ &= \frac{\mathbb{P}(Y_0=i_0,\dots,Y_{m-1}=i_{m-1},Y_m=i,Y_{m+1}=j)}{\mathbb{P}(Y_0=i_0,\dots,Y_{m-1}=i_{m-1},Y_m=i)} \\ &= \frac{\mathbb{P}(X_n=i_0,\dots,X_{m-n+1}=i_{m-1},X_{n-m}=i,X_{n-m-1}=j)}{\mathbb{P}(X_n=i_0,\dots,X_{m-n+1}=i_{m-1},X_{n-m}=i)} \\ &= \frac{\mathbb{P}(X_n=i_0,.,X_{m-n+1}=i_{m-1} \mid X_{n-m}=i,X_{n-m-1}=j)\mathbb{P}(X_{n-m}=i \mid X_{n-m-1}=j)\mathbb{P}(X_{n-m-1}=j)}{\mathbb{P}(X_n=i_0,\dots,X_{m-n+1}=i_{m-1} \mid X_{n-m}=i)\mathbb{P}(X_{n-m}=i)} \\ &= \frac{\mathbb{P}(X_n=i_0,\dots,X_{m-n+1}=i_{m-1} \mid X_{n-m}=i)p(j,i)\pi(j)}{\mathbb{P}(X_n=i_0,\dots,X_{m-n+1}=i_{m-1} \mid X_{n-m}=i)p(i)} \\ &= \frac{p(j,i)\pi(j)}{p(i)}\end{aligned}$$

Thus, $\{Y_m\}$ is a Markov chain with the indicated transition probability.

2.3 Reversibility (Detailed Balance)

Note that reversibility of a Markov process and a reversed Markov process are two entirely different things. There is always a reveresed Markov process, but the fact that it is reversible is a much stronger condition.

Definition 2.14 (Reversibility)

The Markov semigroup $\{P_m\}$ with stationary measure μ is called **reversible** (or in the physics literature, is said to satisfy **detailed balance**) if P_m is self-adjoint for every $f, g, \in L^2(\mu)$. That is,

$$\langle f, P_m g\rangle_\mu = \langle P_m f, g\rangle_\mu$$

By the properties of the adjoint and the Chapman-Kolmogorov equation, we only need to check if P is adjoint.

Note that if the Markov property is reversible, then assuming $X_0 \sim \mu$, then

$$\begin{aligned}\langle P_m f, g\rangle_\mu = \langle f, P_m g\rangle_\mu &= \mathbb{E}[f(X_n)\,\mathbb{E}[g(X_{n+m}) \mid X_n]] \\ &= \mathbb{E}[f(X_n)\,g(X_{n+m})] = \mathbb{E}[\mathbb{E}[f(X_n) \mid X_{n+m})]\,g(X_{n+m}]\end{aligned}$$

for every $f, g \in L^2(\mu)$. So that in particular,

$$P_m f(x) = \mathbb{E}[f(X_{n+m} \mid X_n = x] = \mathbb{E}[f(X_n) \mid X_{n+m} = x]$$

Example 2.10 (Detailed Balance in Finite State Space)

We know that if P is self adjoint, then its transition probability matrix will satisfy

$$\mathbf{P}(i,j) = \frac{\mathbf{P}(j,i)\,\pi(j)}{\pi(i)} \implies \mathbf{P}(j,i)\,\pi(j) = \mathbf{P}(i,j)\,\pi(i)$$

which is the familiar detailed balance condition that we are used to. To see that this is a stronger condition than $\mathbf{P}\boldsymbol{\pi} = \boldsymbol{\pi}$, we sum over j on each side to get

$$\sum_j \mathbf{P}(i,j)\,\pi(i) = \pi(i) \sum_j \mathbf{P}(i,j) = \pi(j)$$

Remember that we could interpret $\pi(i)$ as the amount of water at x, and we send $\mathbf{P}(j,i)\pi(i)$ water from node i to j in one step. The detailed balance condition tells us that the amount of sand going from i to j in one step is exactly balanced by the amount going back from j to i. In contrast, the condition $\boldsymbol{\pi}\mathbf{P} = \boldsymbol{\pi}$ says that after all the transfers are made, the amount of water that ends up at each node is the same as the amount there.

Many chains do not have stationary distributions that satisfy the detailed balance condition.

Example 2.11 ()

Consider the chain with

$$\mathbf{P} = \begin{pmatrix} .5 & .5 & 0 \\ .3 & .1 & .6 \\ .2 & .4 & .4 \end{pmatrix}$$

There is no stationary distribution with detailed balance since $\pi(1)\pi(1,3) = 0$ but $\mathbf{P}(1,3) > 0$ so we must have $\pi(3) = 0$. But this would imply that $\pi(3)\mathbf{P}(3,i) = \pi(i)\mathbf{P}(i,3)$ for all i so we conclude all $\pi(i) = 0$, which doesn't make sense. In fact, the stationary distribution is $(1/3, 1/3, 1/3)$ since $\mathbf{P}$ is doubly stochastic.

2.3.1 Metropolis-Hastings Algorithm

A huge application of Markov chains are in monte carlo algorithms, specifically the Metropolis-Hastings. We begin with a Markov chain with transition probability $q(x,y)$ that is the proposed jump distribution. A move is accepted with probability

$$r(x,y) = \min\left\{\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right\}$$

so the transition probability becomes

$$p(x,y) = q(x,y)r(x,y)$$

Why do we do this? Multiplying by r guarantees that π now satisfies detailed balance under p. Without loss of generality, we can assume $\pi(y)q(y,x) > \pi(x)q(x,y)$, and so we have

$$\begin{aligned} \pi(x)p(x,y) &= \pi(x)q(x,y)\,1 \\ \pi(y)p(y,x) &= \pi(y)q(y,x)\frac{\pi(x)q(x,y)}{\pi(y)q(y,x)} = \pi(x)q(x,y) \end{aligned}$$

which satisfies detailed balance.

2.3.2 Kolmogorov Cycle Condition

Let us take a motivating example.

Example 2.12 ()

Consider the chain with transition probability

$$p = \begin{pmatrix} 1-(a+d) & a & d \\ e & 1-(b+e) & b \\ c & f & 1-(c+f) \end{pmatrix}$$

and suppose that all entries are positive. To satisfy detailed balance, we must have $\pi(x)p(x,y) = \pi(y)p(y,x)$ for all x, y. So we must have

$$e\pi(2) = a\pi(1) \qquad f\pi(3) = b\pi(2) \qquad d\pi(1) = c\pi(3)$$

Multiplying the three equations gives $abc = def$, or in other words,

$$\frac{p(1,2)\,p(2,3)\,p(3,1)}{p(2,1)\,p(3,2)\,p(1,3)} = \frac{abc}{def} = 1$$

Definition 2.15 (Kolmogorov Cycle Condition)

Given a finite irreducible Markov chain with state space S. We say that the **cycle condition** is satisfied if given a cycle of states $x_0, x_1, \dots, x_n = x_0$ with $p(x_{i-1}, x_i) > 0$ for $1 \leq i \leq n$, we have

$$\prod_{i=1}^{n} p(x_{i-1}, x_i) = \prod_{i=1}^{n} p(x_i, x_{i-1})$$

Theorem 2.6 ()

Given a Markov chain S with transition probability p, there exists a stationary distribution π that satisfies detailed balance if and only if the cycle condition holds.

2.4 Ergodicity

Now, we want to talk about "well-behaved" Markov processes that have a limiting distribution that is the stationary measure, i.e. the process will eventually end up in its steady state $\rho_n \to \mu$ as $n \to +\infty$ even if it is not started there. That is, given some fixed initial condition $X_0 = x$, is it true that

$$\mathbb{E}[f(X_n) \mid X_0 = x] \to \mathbb{E}_\mu[f] \text{ as } n \to \infty$$

Definition 2.16 (Ergodicity)

The Markov semigroup (P_n) is called **ergodic** if

$$P_n f \to \mu(f) = \mathbb{E}_\mu[f]$$

as $n \to +\infty$ for every $f \in L^2(\mu)$ (i.e. converges to the constant function $\mu f = \mu(f)$). That is, if we would like to measure $X_n \sim \rho_n$ with f, then far enough in time this measurement converges to measuring $X \sim \mu$ with f. Since this applies to all f (think $f = 1_A$), we can determine that $\rho_n \to \mu$ as $n \to +\infty$.

The following theorem determines whether a chain is ergodic, but note that we don't know anything about the *rate of convergence* to the stationary measure.

Theorem 2.7 ()

If Markov process $\{X_n\}$ with stationary measure μ and semigroup (P_n) is irreducible, then (P_n) is ergodic.

Theorem 2.8 ()

Suppose $|S| < \infty$. If the chain is irreducible and all states positive recurrnent, then there always exists a unique stationary distribution π. If the chain is also aperiodic, then for any initial distribution ν,

$$\lim_{k\to\infty} \nu P^k = \pi$$

Hence

$$\lim_{k\to\infty} P^{(k)}(x,y) = \pi(y)$$

for all $x, y \in S$. Furthermore, for any measurable function $f : S \longrightarrow \mathbb{R}$, the limit

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} f(X_n) = \sum_{x\in S} f(x)\,\pi(x) = \mathbb{E}\big(f(x)\big)$$

holds with probability 1. In particular, the limit does not depend on the initial distribution.

Proof.

The Frobenius Extension to Perron's theorem (Linear Algebra, Theorem 7.31) combined with its applications to stochastic matrices (Linear Algebra, Theorem 7.30) proves this statement.

The next result describes the limiting fraction of time we spend in each state.

Theorem 2.9 (Asymptotic Frequency)

Suppose we have a finite Markov chain with p irreducible and all states recurrent. Then, let

$$N_n(y) = \sum_{i=1}^{n} 1_{X_i = y}$$

be the number of visits to y up to time n. Then,

$$\frac{N_n(y)}{n} \to \frac{1}{\mathbb{E}_y[T_y]}$$

If the chain is aperiodic, then we also have

$$\pi(y) = \frac{1}{\mathbb{E}_y[T_y]}$$

Theorem 2.10 ()

Suppose that a chain is irreducible and there exists stationary distribution π. Then,

$$\frac{1}{n}\sum_{m=1}^{n} p^m(x,y) \to \pi(y)$$

Thus while the sequence $p^m(x,y)$ will not converge in the periodic case, the average of the first n values will.

3 Poisson Processes

3.1 Exponential Distribution

Let us do some review. The **exponential distirbution** of rate λ is a random variable $T \sim \text{Exponential}(\lambda)$ with CDF

$$F_T(t) = \mathbb{P}(T \leq t) = 1 - e^{-\lambda t}$$

and the PDF

$$f_T(t) = \begin{cases} \lambda e^{-\lambda t} & t \geq 0 \\ 0 & t < 0 \end{cases}$$

We have

$$\mathbb{E}[T] = \frac{1}{\lambda}, \quad \text{Var}(T) = \frac{1}{\lambda^2}$$

Lemma 3.1 (Memoryless Property)

The $\text{Exp}(\lambda)$ distribution has the property that for all $t, s \geq 0$,

$$\mathbb{P}(W > t + s \mid W > t) = \mathbb{P}(W > s)$$

which is called the *memoryless property*. We can interpret this in the following way. Let W be the time you have to wait for the first arrival. Given that you already waited t units of time, the probability that you have the wait s additional units of time is just the probability that you wait at least s from the beginning. That is, knowing that t units of time have elapsed does not affect the distribution of the remaining waiting time.

Theorem 3.1 ()

Let W be a continuously distributed random variable. Then $W \sim \text{Exp}(\lambda)$ for some $\lambda > 0$ if and only if W satisfies the memoryless property.

Theorem 3.2 ()

Let $T_i \sim \text{Exponential}(\lambda_i)$ for $i = 1, \dots n$. Then,

$$\min\{T_1, \dots, T_n\} \sim \text{Exponential}(\lambda_1 + \dots + \lambda_n)$$

and the random variable I which takes the index of $\min\{T_1, \dots, T_n\}$ has the PMF

$$\mathbb{P}(I = i) = \frac{\lambda_i}{\lambda_1 + \dots + \lambda_n}$$

3.2 Defining the Poisson Process

We first describe a limiting behavior of binomial random variables.

Theorem 3.3 (Poisson Limit Theorem)

Let $X_n \sim \text{Bernoulli}(n, p_n)$, where $\{p_n\}_{n\in\mathbb{N}}$ is a sequence of reals in $[0,1]$ such that

$$\lim_{n\to\infty} np_n = \lambda$$

Letting $Y \sim \text{Poisson}(\lambda)$

$$X_n \xrightarrow{D} Y$$

That is, the CDFs, and since this is a discrete distribution, the PMFs, converge.

Proof.

We will show that $\lim_{n\to\infty} \mathbb{P}(X_n = k) = \mathbb{P}(Y = k)$, which shows that the CDFs converge and therefore convergence in distribution.

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(X_n = k) &= \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^k \\
&= \lim_{n\to\infty} \frac{n(n-1)\dots(n-k+1)}{k!} \left(\frac{\lambda}{n}\right)^k \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{n^k + O(n^{k-1})}{k!} \frac{\lambda^k}{n^k} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \lim_{n\to\infty} \frac{\lambda^k}{k!} \left(1-\frac{\lambda}{n}\right)^{n-k} \\
&= \frac{\lambda^k}{k!} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{n} \lim_{n\to\infty} \left(1-\frac{\lambda}{n}\right)^{-k} \\
&= \frac{\lambda^k}{k!} e^{\lambda} \, 1 = \frac{\lambda^k e^{\lambda}}{k!}
\end{aligned}$$

Note that this is different from CLT because in CLT, we just assume that the p_n's are constant and take the limiting behavior of $X_n \sim \text{Bernoulli}(n, p)$ as $n \to \infty$.

This result justifies the following model. A Poisson Arrival Process with rate $\lambda > 0$ on the interval $[0, \infty)$ is a model for the occurrence of some events which may have at any time. We can interpret the process as a collection of random points in $[0, \infty)$ which are the times at which the arrivals occur. Suppose that we would like to model the arrival of events that happen completely at random at a rate λ per unit time. At time $t = 0$, we have no arrivals yet, so $N(0) = 0$. Let us fix some T, and now divide $[0, T)$ into n tiny subintervals of length δ.

Assume that in each time slot, we assign a $X_k \sim \text{Bernoulli}(\lambda\delta)$ random variable that determines whether there was an arrival within the interval $((k-1)\delta, k\delta]$. So with probability $\lambda\delta$, there will be an arrival within it, and as the time interval gets smaller, this probability also gets smaller too. Since every n subinterval is $\text{Bernoulli}(\lambda\delta)$, the number of arrivals in the interval $[0, T)$, defined as the random variable $N_n(T)$, is

$$N_n(T) \sim \text{Binomial}(n, \lambda\delta) = \text{Binomial}\big(n, \frac{\lambda T}{n}\big)$$

As we increase the n (equivalently, decrease δ), we divide $[0, T)$ into smaller and smaller subintervals, resulting

in finer and finer $N_n(T)$ Binomial distributions. Since $np_n = n\frac{\lambda T}{n} = \lambda T$ is finite, we can invoke the Poisson limit theorem and say

$$N_n(T) \xrightarrow{D} \text{Poisson}(\lambda T)$$

Note that the starting point 0 does not matter, and this works for any interval of length T. Therefore, we can model the arrival times on any interval of length T as a Poisson(λT) random variable.

Definition 3.1 (Poisson Process)

Let $\lambda > 0$ be fixed, representing the rate of arrival in some unit time. The stochastic counting process $\{N(t)\}_{t \geq 0}$, where $N(t)$ represents the number of arrivals by time t, is called a **Poisson process** with rate λ if

1. $N(0) = 0$
2. The number of arrivals in any interval of length $s > 0$ is $N(t+s) - N(t) \sim \text{Poisson}(\lambda s)$
3. $N(Tt)$ has independent increments, i.e. if $t_0 < t_1 < \ldots, < t_n$, then

$$N(t_1) - N(t_0), \ldots, N(t_n) - N(t_{n-1})$$

are independent.

3.3 Constructing the Poisson Process

Now we have modeled this process using random variables $N(t)$ that counts the number of arrivals up to time t. Now, we can interpret it using random variables that represent the *time* in which they arrive.

Definition 3.2 ()

Set $T_0 = 0$. The arrival times are random variables $0 < T_1 < T_2 < T_3 < \ldots$ such that the inter-arrival waiting times

$$\tau_k = T_k - T_{k-1}, \quad k \geq 0$$

have the property that $\{W_k\}_{k=1}^{\infty}$ are independent Exp(λ) random variables. Define

$$N(s) \coloneqq \max\{k \mid T_k \leq s\}$$

Now we prove that this process is equivalent to the Poisson process defined before.

Theorem 3.4 (Equivalent Interpretations)

Let $\{T_n\}$ be defined as above and $N(s) \coloneqq \max\{k \mid T_k \leq s\}$. Then,

1. $N(0) = 0$
2. $N(s) \sim \text{Poisson}(\lambda s)$
3. $N(t+s) - N(t) \sim \text{Poisson}(\lambda s)$ independent of $N(r)$ for $0 \leq r \leq s$.
4. $N(t)$ has independent incremements.

$N(s) \coloneqq \max\{k \mid T_k \leq s\}$ is a Poisson distribution with mean λs.

4 Continuous-Time Markov Processes

As the name suggests, in a continuous time Markov process X_t, the time parameter is continuous ($t \geq 0$). As before, the system jumps randomly between states in S, but now the jumps may occur at any time and they occur randomly. This implies that there are *two* sources of randomness:

1. *where* the system jumps, which is determined by the transition probabilities, and
2. *when* the system jumps, which is called the holding time

Definition 4.1 (CTMP)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ a measurable space. Then, a homogeneous **continuous-time Markov chain** is a stochastic process $\{X_t\}_{t\geq 0}$ taking values in S (i.e. $X_t : \Omega \to S$) satisfying the **Markov property**: for every bounded measurable f and and $t, s \geq 0$,

$$\mathbb{E}[f(X_{t+s}) \mid \{X_r\}_{r\leq t}] = \mathbb{E}[f(X_{t+s}) \mid X_t] = (P_s f)(X_t)$$

This again says that the probability of X_{t+s} does not depend on the history $\{X_r = i_r\}_{r\leq t}$, but on the current value of X_t.

Just like the discrete-time case, to describe random variable X_{t+s} with function f, we can pull back the function to compute

$$\mathbb{E}[f(X_{t+s})] = \mathbb{E}[\mathbb{E}[f(X_{t+s}) \mid X_t]] = \mathbb{E}[(P_s f)(X_t)] = \int_S P_s f \, d\rho_t$$

which integrates a new function $P_s f$ over the measure ρ_t.

Example 4.1 (Transition Operator as a Matrix in Discrete Space)

Let us have a discrete space $S = \{1, \dots, d\}$ with indicators $1_{\{i\}}$ for $i = 1, \dots, d$. Let x_t represent the column vector of the PMF of X_t. From the same work as shown for discrete time Markov processes, we can let $f = 1_{\{j\}}$ and compute the probability of X_{t+s} landing in each point $j \in S$, since that is what we're interested in for discrete probability distributions.

$$\begin{aligned}
\boldsymbol{\rho_{t+s}}(j) &= \mathbb{P}(X_{t+s} = j) \\
&= \mathbb{E}[1_{\{j\}}(X_{t+s})] \\
&= \mathbb{E}[\mathbb{E}[1_{\{j\}}(X_{t+s}) \mid X_t]] && = \mathbb{E}[P_s 1_{\{j\}}(X_t)] \\
&= \int_S \mathbb{E}[1_{\{j\}}(X_{t+s}) \mid X_t] d\rho_t && = \int_S P_s 1_{\{j\}}(X_t) \, d\rho_t \\
&= \sum_{i\in S} \mathbb{P}[X_{t+s} = j \mid X_t = i] \, \mathbb{P}(X_t = i) && = \sum_{i\in S} P_s 1_{\{j\}}(i) \mathbb{P}(X_t = i)
\end{aligned}$$

which can be summarized as

$$\boldsymbol{\rho_{t+s}}(j) = \sum_{i=1}^{d} P_s 1_{\{j\}}(i) \boldsymbol{\rho_t}(i) = \sum_{i=1}^{d} \mathbb{P}(X_{t+s} = j \mid X_t = i) \, \boldsymbol{\rho_t}(i)$$

We can compactly organize the probabilities of these internode travel inside a $d \times d$ right stochastic **transition matrix**

$$\mathbf{P_s} = \begin{pmatrix} P_s 1_{\{1\}}(1) & \dots & P_s 1_{\{1\}}(d) \\ \vdots & \ddots & \vdots \\ P_s 1_{\{d\}}(1) & \dots & P_s 1_{\{d\}}(d) \end{pmatrix} = \begin{pmatrix} \mathbb{P}(X_{t+s} = 1 \mid X_t = 1) & \dots & \mathbb{P}(X_{t+s} = d \mid X_t = 1) \\ \vdots & \ddots & \vdots \\ \mathbb{P}(X_{t+s} = 1 \mid X_t = d) & \dots & \mathbb{P}(X_{t+s} = d \mid X_t = d) \end{pmatrix}$$

and compactly write the above equation as

$$\boldsymbol{\rho}_{\boldsymbol{t+s}}^T = \boldsymbol{\rho}_{\boldsymbol{t}}^T \mathbf{P_s}$$

Lemma 4.1 ()

P_t is linear. That is, for $t, s \geq 1$, $\alpha, \beta \in \mathbb{R}$, and bounded measurable functions f, g,

$$P_t(\alpha f + \beta g) = \alpha P_t f + \beta P_t g$$

Proof.

By linearity of conditional expectation,

$$\begin{aligned}(P_s(\alpha f + \beta g))(X_t) &= \mathbb{E}[(\alpha f + \beta g)(X_{t+s}) \mid X_t] \\ &= \mathbb{E}[(\alpha f)(X_{t+s}) \mid X_t] + \mathbb{E}[(\beta g)(X_{t+s}) \mid X_t] \\ &= \alpha(P_s f)(X_t) + \beta(P_s g)(X_t)\end{aligned}$$

We can now interpret linearity and the Markov property in the discrete space.

Example 4.2 (Markov Property in Discrete Space)

If we wanted to extract information from X_t with function f (i.e. compute $\mathbb{E}[f(X_t)]$), we can calculate

$$\mathbb{E}[f(X_t)] = \boldsymbol{\rho_t}^T \mathbf{f} = \begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

Now, say that s units of time later, we want to extract information f from X_{t+s} by computing

$$\mathbb{E}[f(X_{t+s})] = \boldsymbol{\rho_{t+s}}^T \mathbf{f} = \begin{pmatrix}\boldsymbol{\rho_{t+s}}(1) & \dots & \boldsymbol{\rho_{t+s}}(d)\end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

The problem is that we don't know what the distribution of X_{t+s} is (i.e. don't know $\boldsymbol{\rho_{t+s}}(i)$), so we get its expectation by conditioning it on X_t, which realizes as taking the expectation of a *different* function $P_s f$ with respect to ρ_t.

$$\mathbb{E}[f(X_{t+s})] = \mathbb{E}[\mathbb{E}[f(X_{t+s}) \mid X_t]] = \mathbb{E}[(P_s f)(X_t)] = \begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \begin{pmatrix} (P_s f)_1 \\ \vdots \\ (P_s f)_d \end{pmatrix}$$

It turns out that this transformation $\mathbf{f} \mapsto \mathbf{P_s f}$ (from row vector to row vector) is linear, and so we can interpret $\mathbf{P_s}$ as $\mathbf{f}$ that has been left-multiplied by some transformation matrix $\mathbf{P_s}$.

$$\begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \begin{pmatrix} (P_s f)_1 \\ \vdots \\ (P_s f)_d \end{pmatrix} = \begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \underbrace{\begin{pmatrix} & \mathbf{P_s} & \end{pmatrix} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}}_{\mathbf{P_s f}}$$

It turns out that this $\mathbf{P_s}$ acts linearly on $\mathbf{f}$ through left multiplication, but we can also right-multiply $\boldsymbol{\rho_t}$ by $\mathbf{P_s}$ to get the new distribution of X_{t+s}!

$$\begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \begin{pmatrix} (P_s f)_1 \\ \vdots \\ (P_s f)_d \end{pmatrix} = \underbrace{\begin{pmatrix}\boldsymbol{\rho_t}(1) & \dots & \boldsymbol{\rho_t}(d)\end{pmatrix} \begin{pmatrix} & \mathbf{P_s} & \end{pmatrix}}_{\boldsymbol{\rho_t}^T \mathbf{P_s} = \boldsymbol{\rho}_{t+s}^T} \begin{pmatrix} f_1 \\ \vdots \\ f_d \end{pmatrix}$$

Therefore, it turns out that the linearity of $\mathbf{P_s}$ on $\mathbf{f}$ implies linearity of it on the vector $\boldsymbol{\rho_t}$.

Now focusing on $f = 1_A$, we can define the following.

Definition 4.2 (Transition Probability)

Let us have Markov process (X_t) with operator P_s. The function $p_s : S \times \mathcal{S} \to \mathbb{R}$ defined

$$p_s(x, A) \coloneqq P_s 1_A(x) = \mathbb{E}[1_A(X_{t+s}) \mid X_t = x] = \mathbb{P}(X_{t+s} \in A \mid X_t = x)$$

is the **transition probability**, or **transition kernel**, of this chain. Note that

1. For each $x \in S$, $A \mapsto p_s(x, A)$ is a probability measure on $(S, \mathcal{S})$. This means that if we are in some place x at time t, then the probability that we will land in some subset $A \in \mathcal{S}$ of S at time $t + s$ is $p_s(x, A)$.
2. For each $A \in \mathcal{S}$, $P_s 1_A = p_s(\cdot, A)$ is a measurable function.

The **transition kernel density** is simply the pdf of the measure $p_s(x, \cdot)$.

$$p_s(x, A) = \int_A p_s(x, y)\, dy$$

Note that in the matrix realization of the example above, it looks like P_s acts on the distribution ρ_t to get a new distribution ρ_{t+s}, but this is not strictly the case since P_s is an operator on f. However, for the sake of intuitiveness, we can interpret P_s in two ways:

1. It operates on the measure ρ_t by pushing it forward in time to get ρ_{t+s}. This operator is defined as

   $$\rho_t \mapsto \rho_{t+s}(\cdot) = p_s(X_t, \cdot)$$

   which corresponds to the matrix multiplication $\boldsymbol{\rho_t^T} \mapsto \boldsymbol{\rho_{t+s}^T} = \boldsymbol{\rho_t^T \mathbf{P_s}}$
2. It operates on the function f (at X_{t+s}) by pulling it back to $P_s f$ that operates on X_t. This operation $f \mapsto P_s f$ corresponds to the matrix multiplication $\mathbf{f} \mapsto \mathbf{P_s f}$.

Either way, we can think of the order of operations as either $(\boldsymbol{\rho_t^T \mathbf{P_s}})\mathbf{f}$ or $\boldsymbol{\rho_t^T}(\mathbf{P_s f})$.

Just like stochastic transition matrices, we can also deduce a semigroup property of the collection $(P_s)_{s \geq 0}$.

Lemma 4.2 (Chapman-Kolmogorov)

$\{P_t\}$ satisfies

$$P_{t+s} f = P_t P_s f$$

for all $t, s, \geq 1$, with $P_0 = I$, the identity.

Proof.

We can easily see that $(P_0 f)(X_t) = \mathbb{E}[f(X_t) \mid X_t] = f(X_t)$, and

$$\begin{aligned}
(P_{t+s} f)(X_n) &= \mathbb{E}[f(X_{n+t+s}) \mid X_n] \\
&= \mathbb{E}[\mathbb{E}[f(X_{n+t+s} \mid X_{n+t})] \mid X_n] \\
&= \mathbb{E}[(P_s f)(X_{n+t}) \mid X_n] \\
&= (P_t(P_s f))(X_n) \\
&= (P_t P_s f)(X_n)
\end{aligned}$$

We give one final condition.

Lemma 4.3 (Conservativeness)

$\{P_t\}$ satisfies

$$P_t 1 = 1$$

for all $t \geq 0$, where $1 = 1_S$ is the constant function of 1.

Proof.

This is trivial since it is just the law of total probability. That is, $1_S(X_t) = 1$, and

$$(P_s 1_S)(X_t) = \mathbb{E}[1_S(X_{t+s}) \mid X_t]$$

and note that $\sigma(X_t)$ is a finer σ-algebra than that generated by $1_S(X_{t+s})$, meaning that the right hand side is equal to $1_S(X_{t+s})$ itself, which equals 1.

Example 4.3 ()

Given the transition matrix

$$\mathbf{P_s} = \begin{pmatrix} P_s 1_{\{1\}}(1) & \dots & P_s 1_{\{1\}}(d) \\ \vdots & \ddots & \vdots \\ P_s 1_{\{d\}}(1) & \dots & P_s 1_{\{d\}}(d) \end{pmatrix}$$

note that by linearity of P_s and the fact that $\{j\}$ forms a partition of S, we have a

$$\sum_{j \in S} (P_s 1_{\{j\}})(i) = \left[P_s \left(\sum_{j \in S} 1_{\{j\}} \right) \right](i) = (P_s 1_S)(i) = 1_S(i) = 1$$

which means that the columns must sum to 1.

Example 4.4 (Markov Chain with Continuous Jumps)

Let $N(t), t \geq 0$ be a Poisson process with rate λ and let Y_n be a discrete time Markov chain with transition probability $u(i, j)$. Then, $X_t = Y_{N(t)}$ is a continuous time Markov chain that takes one jump according to $u(i, j)$ at each arrival time $N(t)$.

4.1 Generator

In the discrete time case, we had $P_t = (p_1)^t$ for $t \in \mathbb{N}$, and from the Chapman-Kolmogorov equation, knowing p_1 allows us to compute p_t for all $t \in \mathbb{N}$. Likewise, if we know the transition probability for some $t < t_0$ for any $t_0 > 0$, we know it for all t. This observation suggests that the transition probabilities p_t can be determined from their derivatives at 0.

We now define the analogous operator to the transition rate matrix in continuous-time chains with a finite state space. This is a natural extension, since we are just taking the right-derivative of P_t at $t = 0$.

Definition 4.3 (Generator)

The generator $\mathscr{L}$ is defined as

$$\mathscr{L} f := \lim_{t \downarrow 0} \frac{P_t f - f}{t}$$

for every $f \in L^2(\mu)$ for which the above limit exists in $L^2(\mu)$. Intuitively, $\mathscr{L} f$ represents the instan-

taneous rate of change of the measurement f. The set of f for which $\mathscr{L}f$ is defined is called the domain $\text{Dom}(\mathscr{L})$ of the generator, and $\mathscr{L}$ defines a linear operator from $\text{Dom}(\mathscr{L}) \subset L^2(\mu)$ to $L^2(\mu)$.

We have defined the generator $\mathscr{L}$ from the Markov semigroup $\{P_t\}_{t\geq 0}$. Now, let's try to define the semigroup in terms of the generator $\mathscr{L}$. Given that we have some map $\mathscr{L}$), can we define some semigroup $\{P_t\}$ satisfying the definition? We know that by the semigroup property, we can split P_{t+h} into P_tP_h and P_hP_t, from which we get the **Kolmogorov backward equation** and the **forward equation**, respectively.

$$\frac{d}{dt}P_t = \lim_{h\downarrow 0}\frac{P_{t+h}-P_t}{h} = \lim_{h\downarrow 0}\frac{P_t(P_h-I)}{h} = P_t\left(\lim_{h\downarrow 0}\frac{P_h-I}{h}\right) = P_t\mathscr{L}$$
$$\frac{d}{dt}P_t = \lim_{h\downarrow 0}\frac{P_{t+h}-P_t}{h} = \lim_{h\downarrow 0}\frac{(P_h-I)P_t}{h} = \left(\lim_{h\downarrow 0}\frac{P_h-I}{h}\right)P_t = \mathscr{L}P_t$$

From which we see that the generator $\mathscr{L}$ is commutes with the semigroup

$$\mathscr{L}P_t = P_t\mathscr{L}$$

and solving this differential equation gives

$$P_t = e^{t\mathscr{L}}$$

Let's observe how this generator acts on the indicator functions $f = 1_A$. Note that $P_s1_A(i) = \mathbb{P}(X_{t+s}\in A \mid X_t = i)$.

$$(\mathscr{L}1_A)(i) = \left(\lim_{h\downarrow 0}\frac{P_h1_A - 1_A}{h}\right)(i) = \lim_{h\downarrow 0}\frac{P_h1_A(i)-1_A(i)}{h}$$

and so $(\mathscr{L}1_A)(i)$ represents the infinitesimal rate of change of the probability that X_t will be in A given that it is at 1.

Now, how does the generator realize into the finite state space?

Example 4.5 (Transition Rate Matrix)

We know that the semigroup operator P_t is equivalent to the transition matrix

$$\mathbf{P_t} = \begin{pmatrix} P_t(1,1) & \dots & P_t(1,d) \\ \vdots & \ddots & \vdots \\ P_t(d,1) & \dots & P_t(d,d) \end{pmatrix}$$

Let's say that we have the function $f = \sum_{i\in S} c_i 1_{\{i\}}$, which realizes as the function vector $\mathbf{f}$, and we have generator $\mathscr{L}$. We know that P_tf realizes as the matrix multiplication $\mathbf{P_tf}$, and so we can define the **transition rate matrix Q** satisfying the equation

$$\mathbf{Qf} = \lim_{h\to 0}\frac{\mathbf{P_hf}-\mathbf{f}}{h} \implies \mathbf{Q} = \lim_{h\to 0}\frac{\mathbf{P_h}-\mathbf{I}}{h}$$

This derivatives has entries

$$Q(i,j) = \frac{d}{dt}\bigg|_{t=0}\mathbf{P_t}(i,j) = \lim_{h\to 0}\frac{\mathbf{P_h}(i,j)-\mathbf{P_0}(i,j)}{h} = \begin{cases} \lim_{h\to 0}\frac{P_h(i,j)}{h} & \text{if } i\neq j \\ \lim_{h\to 0}\frac{P_h(i,i)-1}{h} & \text{if } i = j \end{cases}$$

representing the flow of probability from $i \mapsto j$. Note that by the law of total probability,

$$\sum_j \mathbf{P_t}(i,j) = 1 \implies \frac{d}{dt}\bigg|_{t=0}\sum_j \mathbf{P_t}(i,j) = \sum_j \frac{d}{dt}\bigg|_{t=0}\mathbf{P_t}(i,j) = \sum_j \mathbf{Q}(i,j) = 0$$

So the diagonal entries is simply $\mathbf{Q}(i,i) = -\sum_{j\neq i} Q(i,j)$. This realization $\mathbf{Q}$ is consistent with the way $\mathscr{L}$ operates. Given $f = \sum_i f_i 1_{\{i\}}$, and not worrying about whether we evaluate a limit of functions or the limit of evaluations, we can get

$$\begin{aligned}(\mathscr{L}f)(i) &= \left[\mathscr{L}\left(\sum_{j=1}^d f_j 1_{\{j\}}\right)\right](i) = \left(\sum_{j=1}^d f_j \mathscr{L}1_{\{j\}}\right)(i) = \sum_{j=1}^d f_j(\mathscr{L}1_{\{j\}})(i) \\ &= \sum_{j=1}^d f_j\left(\lim_{h\downarrow 0}\frac{P_h 1_{\{j\}}(i) - 1_{\{j\}}(i)}{h}\right) = \sum_{j=1}^d f_j\left(\lim_{h\downarrow 0}\frac{\mathbf{P_h}(i,j) - \mathbf{P_0}(i,j)}{h}\right) \\ &= \sum_{j=1}^d \mathbf{Q}(i,j) f_j = (\mathbf{Qf})_i\end{aligned}$$

and therefore, setting $f = 1_{\{j\}}$, we get

$$\mathscr{L}1_{\{j\}}(i) = Q(j,i)$$

Example 4.6 ()

Given a two-state Markov chain, $\{0,1\}$, with some $\lambda \geq 0$. Then, we can model our transition probability matrix as

$$P_s = \begin{pmatrix} \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} \\ \frac{1}{2} - \frac{1}{2}e^{-2\lambda t} & \frac{1}{2} + \frac{1}{2}e^{-2\lambda t} \end{pmatrix}$$

Its generator matrix is

$$Q = \begin{pmatrix} -\lambda & \lambda \\ \lambda & -\lambda \end{pmatrix}$$

4.2 Classification of States

4.2.1 Holding Times and Jumping Times

Now, we would like to find how long a chain stays at a state $x \in S$.

Definition 4.4 (Holding Time)

Let $\{X_t\}_{t\geq 0}$ be a continuous time Markov chain, and define T_x to be the **holding time** at x.

$$X_t = x, \quad T_x = \inf\{s \geq t, X_s \neq x\}$$

We can characterize the distribution of T_x, but first we define the following.

Definition 4.5 (Memoryless Property)

A random variable X has the **memoryless property** if it satisfies for all $t, s \geq 0$

$$\mathbb{P}(X > s + t \mid X > t) = \mathbb{P}(X > s)$$

which is just abuse of notation for the following: We know that (t,∞), (s,∞), and $(s+t,\infty)$ are all in $\mathcal{R}$ and so they are events. So it really translates to the probability of an outcome landing in $(s+t,\infty)$ given that it lands in (t,∞) is equal the probability of it landing in (s,∞).

$$\mathbb{P}_X\big((s+t,\infty) \mid (t,\infty)\big) = \frac{\mathbb{P}_X\big((s+t,\infty)\cap(t,\infty)\big)}{\mathbb{P}_X\big((t,\infty)\big)} = \frac{\mathbb{P}_X\big((s+t,\infty)\big)}{\mathbb{P}_X\big((t,\infty)\big)} = \mathbb{P}_X\big((s,\infty)\big)$$

The exponential random variable is memoryless because the LHS just reduces to

$$\frac{\mathbb{P}_X\big((s+t,\infty)\big)}{\mathbb{P}_X\big((t,\infty)\big)} = \frac{1-F_X(s+t)}{1-F_X(t)} = \frac{e^{-\lambda(s+t)}}{e^{-\lambda t}} = e^{-\lambda s} = 1-F_X(s) = \mathbb{P}_X\big((s,\infty)\big)$$

Theorem 4.1 ()

The only continuous random variable having the memoryless property is the exponential random variable.

Theorem 4.2 ()

T_x has the memoryless property.

Proof.

We can show that

$$\begin{aligned}\mathbb{P}(T_x > t+s \mid T_x > t) &= \mathbb{P}(X_u = x,\, u\in[t,t+s] \mid X_u = x,\, u\in[0,t]) \\ &= \mathbb{P}(X_u = x,\, u\in[t,t+s] \mid X_t = x) \\ &= \mathbb{P}(T_x > s)\end{aligned}$$

Therefore, we know that T_x must have the exponential distribution, and for each x, we have $T_x \sim \text{Exp}(\lambda_x)$.

4.2.2 Irreducibility

Definition 4.6 (Irreducibility)

The Markov chain X_t is **irreducible** if for any two states $i, j \in S$, it is possible to get from i to j in a finite number of steps. To be precise, there is a sequence of states $k_0 = i, k_1, \dots, k_n = j$ s.t.

$$Q(k_{m-1}, k_m) > 0$$

Lemma 4.4 ()

If X_t is irreducible and $t > 0$, then $P_t(i,j) > 0$ for all $i, j \in S$.

4.3 Stationary Measures

Recall that the Markov process $(X_t)_{t\geq 0}$ evolves, and this evolution can be described by the sequence of measures $(\rho_t)_{t\geq 0}$ for each X_t. If we would like to measure X_{t+s} with function f, we can calculate $\mathbb{E}[f(X_{t+s})] = \mathbb{E}_{\rho_{t+s}}[f]$, but we don't know ρ_{t+s}. Fortunately, we can "pull back" the f to compute the equivalent

$$\mathbb{E}_{\rho_{t+s}}[f] = \mathbb{E}[f(X_{t+s})] = \mathbb{E}[\mathbb{E}[f(X_{t+s}) \mid X_t]] = \mathbb{E}[P_s f(X_t)] = \mathbb{E}_{\rho_t}[P_s f]$$

which essentially measures X_{t+s} with f by measuring X_t with $P_s f$. Now, we want to construct a stationary measure that captures the fact that if a certain state $X_t \sim \rho_t = \mu$ follows a stationary measure, then the measure of future $X_{t+s} \sim \rho_{t+s} = \mu$ also. If μ is stationary, then both $\rho_{t+s} = \rho_t = \mu$, and this is equivalent to

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_s f]$$

for all measure f and $s \geq 0$. This will be the definition that we will work with. To help with the interpretation, we can restrict the case to $f = 1_A$ to get $\mathbb{P}(X_t \in A) = \mathbb{P}(X_{t+s} \in A)$ for all $A \in \mathcal{S}$, which means that the probability of X_{t+s} realizing in A is equal to the probability of X_t realizing in A. In summary, stationary measures describe the equilibrium or steady-state behavior of the Markov process.

Definition 4.7 (Stationary Measure)

A probability measure μ is called **stationary** or **invariant** if

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_t f], \text{ conventionally written as } \mu(f) = \mu(P_t f)$$

for all $t \geq 0$ and bounded measurable f. This is a property of the *measure*. We can describe the way it operates on the measure as if $\rho_t = \mu$, then

$$\rho_{t+s}(\cdot) = p_s(X_t, \cdot) = \rho_t$$

To give a pictorial interpretation, imagine an initial distribution $X_0 \sim \rho_0$ as some amount of sand placed on the state space S (either as a continuous mass or mounds on discrete nodes). As time flows continuously, the distribution will evolve to $X_t \sim \rho_t$, where a different mound of sand will form on S. If $\rho_0 = \mu$, then the flow of sand between the nodes will balance each other out, and we still have the same amount of sand $\rho_t = \mu$ after each step. The discrete case is simpler, since we can just imagine there being $\boldsymbol{\pi}(i)$ of sand at node i, and $\mathbf{P_t}(i,j)$ of its proportion of sand flowing from node i to j after time t. Therefore, all the sand flowing out of i, which is $\sum_{j=1}^d \mathbf{P_t}(i,j)\boldsymbol{\pi}(i) = 1$, balances out with the flow of sand into i, which is $\sum_{j=1}^d P(j,i)\boldsymbol{\pi}(j)$.

$$1 = \sum_{i=1}^d P(i,j)\boldsymbol{\pi}(i) = \sum_{j=1}^d P(j,i)\boldsymbol{\pi}(j)$$

and doing this for all i realizes into the matrix equation $\boldsymbol{\pi} = \boldsymbol{\pi}\mathbf{P_t}$.

Example 4.7 (Stationary Distribution in Discrete Space)

Given discrete state space $S = \{1, \dots, d\}$, our stationary measure μ can be represented by the all familiar row vector

$$\boldsymbol{\pi} = \begin{pmatrix} \boldsymbol{\pi}(1) & \dots & \boldsymbol{\pi}(d) \end{pmatrix} = \begin{pmatrix} \mu(\{1\}) & \dots & \mu(\{d\}) \end{pmatrix}$$

Given the PMF vectors $\boldsymbol{\rho_t} = \boldsymbol{\pi}$ and $\boldsymbol{\rho_{t+s}} = \boldsymbol{\pi}$ and some measurable function $\mathbf{f} = (f_1, \dots, f_d)$, the stationary distribution property says that

$$\mathbb{E}[f(X_{n+m})] = \mathbb{E}[(P_m s f)(X_n)] \iff \boldsymbol{\pi}\mathbf{f} = \boldsymbol{\pi}\mathbf{P_m}\mathbf{f}$$

which means that $\mathbf{P_s}\mathbf{f}$ will act on $\boldsymbol{\pi}$ the same way that $\mathbf{f}$ does (though $\mathbf{P_s}\mathbf{f} \neq \mathbf{f}$). We can also interpret $\boldsymbol{\pi}$ as the eigenvector of $\mathbf{P_s}$ with eigenvalue 1 since $\rho_{t+s}(\cdot) = p_s(X_t, \cdot) = \rho_t(\cdot)$.

Theorem 4.3 ()

If μ is a stationary measure of a continuous-time Markov process with generator $\mathscr{L}$, then

$$\mu(\mathscr{L}f) = 0$$

for every $f \in L^2(\mu)$.

Proof.

Not worrying about interchanging limits and integrals, we have

$$\begin{aligned}\mu(\mathscr{L}f) = \mathbb{E}_\mu[\mathscr{L}f] &= \int_S \lim_{t\downarrow 0} \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \int_S \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \frac{1}{t}\big(\mathbb{E}_\mu[P_t f] - \mathbb{E}_\mu[f]\big) = \lim_{t\downarrow 0} \frac{1}{t}\cdot 0 = 0\end{aligned}$$

For a finite state space, this theorem reduces to the following.

Corollary 4.1 ()

π is a stationary distribution of a continuous time Markov chain if and only if

$$\boldsymbol{\pi Q = 0}$$

Proof.

To prove the if, we have

$$\pi Q = 0 \implies \pi P_t = \pi e^{tQ} = \pi\left(I + tQ + \frac{t^2Q^2}{2!} + \dots\right) = \pi + 0 + \dots = \pi$$

To prove the only if, we have

$$\pi P_t = \pi \implies 0 = \frac{d}{dt}\pi P_t = \pi \frac{d}{dt} P_t = \pi Q P_t \implies \pi Q = 0$$

Theorem 4.4 ()

If a continuous-time Markov chain X_t is irreducible and has a stationary distribution π, then

$$\lim_{t\to\infty} P_t(i,j) = \pi(j)$$

4.3.1 Uniqueness

TBD TBD

4.3.2 Reversed Markov Process

From now, given the state space $(S, \mathcal{S})$ we can put a measure μ on it to get a measure space $(S, \mathcal{S}, \mu)$. The Banach space of all μ-measurable functions $f : (S, \mathcal{S}, \mu) \to (\mathbb{R}, \mathcal{R})$ (i.e. for every Borel $B \in \mathcal{R}$, $f^{-1}(B) \in \mathcal{S}$) will be denoted $L^p(\mu)$, equipped with the norm

$$||f||_{L^p(\mu)} \coloneqq \mathbb{E}_\mu[f^p]^{1/p} = \left(\int_S |f|^p \, d\mu \right)^{1/p}$$

If $p = 2$, then we can define the inner product

$$\langle f, g \rangle_\mu \coloneqq \mathbb{E}_\mu[fg] = \int_S fg \, d\mu$$

Lemma 4.5 (Contraction of Stationary Measure)

Let μ be a stationary measure. Then,

$$||P_t f||_{L^p(\mu)} \geq ||f||_{L^p(\mu)} = \mathbb{E}_\mu[f^p]^{1/p}$$

Now, we can construct reversed Markov processes.

Definition 4.8 (Reversed Markov Process)

Let $\{X_t\}_{0 \leq t \leq T}$ be a continuous time Markov process with semigroup $(P_t)_{t \geq 0}$ and stationary distribution μ. Then, fix T and let $Y_t = X_{T-t}$. Then, Y_t is a discrete time Markov process with the **dual transition operator** P_t^*, the adjoint of P_t satisfying

$$\langle f, P_t g \rangle_\mu = \langle P_t^* f, g \rangle_\mu$$

for all bounded measurable $f, g \in L^2(\mu)$.

Though we have given the reversed Markov process as a definition above, we can prove that this satisfies the Markov property.

Proof.

We can see how this definition realizes in a discrete space.

Example 4.8 ()

Given $S = \{1, \dots, d\}$ and function vectors $\mathbf{f}, \mathbf{g}$,

$$\langle f, g \rangle_\mu = \int_S fg d\mu = \sum_{i=1}^d f_i g_i \pi(i)$$

and by definition of the adjoint, we must have

$$\begin{aligned} \langle f, P_t g \rangle_\mu = \sum_{i=1}^d f_i (\mathbf{P_t g})_i \pi(i) &= \sum_{i=1}^d f_i \left(\sum_{j=1}^d \mathbf{P_t}(i,j) g_j \right) \pi(i) \\ &= \sum_{i=1}^d g_i \left(\sum_{j=1}^d \mathbf{P_t^*}(i,j) f_j \right) \pi(i) = \sum_{i=1}^d (\mathbf{P_t^* f})_i \, g_i \, \pi(i) = \langle P_t^* f, g \rangle_\mu \end{aligned}$$

A bit of computation will show us that

$$\mathbf{P_t^*}(i,j) = \frac{\mathbf{P_t}(j,i)\pi(j)}{\pi(i)}$$

and we can indeed check that

$$\begin{aligned}
\langle P_t^* f, g\rangle_\mu &= \sum_{i=1}^d g_i \left(\sum_{j=1}^d \mathbf{P}_\mathbf{t}^*(i,j) f_j \right) \pi(i) \\
&= \sum_{i=1}^d g_i \left(\sum_{j=1}^d f_j \frac{\mathbf{P}_\mathbf{t}(j,i)\pi(j)}{\pi(i)} \right) \pi(i) \\
&= \sum_{j=1}^d \sum_{i=1}^d g_i \, f_j \mathbf{P}_\mathbf{t}(j,i) \, \pi(j) \\
&= \sum_{j=1}^d f_j \left(\sum_{i=1}^d g_i \mathbf{P}_\mathbf{t}(j,i) \right) \pi(j) \\
&= \sum_{j=1}^d f_j (\mathbf{P_t g})_j \pi(j) = \langle f, P_t g\rangle_\mu
\end{aligned}$$

Note that $\mathbf{P}_\mathbf{t}^*$ also satisfies $\mathbf{P}_\mathbf{t}^*(i,j) \geq 0$ and by definition of the stationary distribution π,

$$\sum_{j=1}^d \mathbf{P}_\mathbf{t}^*(i,j) = \sum_{j=1}^d \frac{\mathbf{P}_\mathbf{t}(j,i)\pi(j)}{\pi(i)} = \frac{1}{\pi(i)} \sum_{j=1}^d \mathbf{P}_\mathbf{t}(j,i)\pi(j) = \frac{\pi(i)}{\pi(i)} = 1$$

Note that the transition probability is computed using Bayes rule

$$\begin{aligned}
\mathbf{P}_\mathbf{s}^*(i,j) &= \mathbb{P}(Y_{t+s} = j \mid Y_t = i) \\
&= \frac{\mathbb{P}(Y_t = i \mid Y_{t+s} = j)\mathbb{P}(Y_{t+s} = j)}{\mathbb{P}(Y_t = i)} \\
&= \frac{\mathbb{P}(X_{T-t} = i \mid X_{T-t-s} = j)\mathbb{P}(X_{T-t-s} = j)}{\mathbb{P}(X_{T-t} = i)} \\
&= \frac{\mathbf{P}_s(j,i)\pi(j)}{\pi(i)}
\end{aligned}$$

4.4 Reversibility (Detailed Balance)

Note that reversibility of a Markov process and a reversed Markov process are two entirely different things. There is always a reveresed Markov process, but the fact that it is reversible is a much stronger condition.

Definition 4.9 (Reversibility)

The Markov semigroup $\{P_s\}$ with stationary measure μ is called **reversible** (or in the physics literature, said to satify **detailed balance**) if P_s is self-adjoint for every $f, g, \in L^2(\mu)$. That is,

$$\langle f, P_s g\rangle_\mu = \langle P_s f, g\rangle_\mu$$

Since $P_s = e^{s\mathscr{L}}$, this condition is equivalent to $\mathscr{L}$ being self-adjoint.

Note that if the Markov property is reversible, then assuming $X_0 \sim \mu$, then

$$\begin{aligned}\langle P_s f, g\rangle_\mu = \langle f, P_s g\rangle_\mu &= \mathbb{E}[f(X_t)\,\mathbb{E}[g(X_{t+s}) \mid X_t]] \\ &= \mathbb{E}[f(X_t)\,g(X_{t+s})] = \mathbb{E}[\mathbb{E}[f(X_t) \mid X_{t+s})]\,g(X_{t+s}]\end{aligned}$$

for every $f, g \in L^2(\mu)$. So that in particular,

$$P_s f(x) = \mathbb{E}[f(X_{t+s} \mid X_t = x] = \mathbb{E}[f(X_t) \mid X_{t+s} = x]$$

which means that the reversed process follows the same law as the forwrad process.

Example 4.9 (Detailed Balance in Finite State Space)

We know that if P_s is self adjoint, then its transition probability matrix will satisfy

$$\mathbf{P_s}(i,j) = \frac{\mathbf{P_s}(j,i)\,\pi(j)}{\pi(i)} \implies \mathbf{P_s}(j,i)\,\pi(j) = \mathbf{P_s}(i,j)\,\pi(i)$$

which is the familiar detailed balance condition that we are used to. To see that this is a stronger condition than $\boldsymbol{\pi}\mathbf{P_t} = \boldsymbol{\pi}$, we sum over j on each side to get

$$\sum_j \mathbf{P_s}(i,j)\,\pi(i) = \pi(i) \sum_j \mathbf{P_s}(i,j) = \pi(j)$$

Remember that we could interpret $\pi(i)$ as the amount of water at x, and we send $\mathbf{P_s}(j,i)\pi(i)$ water from node i to j in one step. The detailed balance condition tells us that the amount of sand going from i to j in one step is exactly balanced by the amount going back from j to i. In contrast, the condition $\boldsymbol{\pi}\mathbf{P_s} = \boldsymbol{\pi}$ says that after all the transfers are made, the amount of water that ends up at each node is the same as the amount there.

4.5 Ergodicity

Now, given a Markov semigroup P_t with generator $\mathscr{L}$ and stationary measure μ, we know that $X_0 \sim \mu$ implies $X_t \sim \mu$ for all times t. It is natural to ask whether the Markov process will eventually end up in its steady state even if it is not started there, but rather at some fixed initial condition. That is, given $X_0 = x$, is it true that

$$\mathbb{E}[f(X_t) \mid X_0 = x] \to \mu f = \mathbb{E}_\mu[f] \text{ as } t \to \infty$$

If this is the case, the Markov process is said to be ergodic.

Definition 4.10 (Ergodicity)

The Markov semigroup (P_t) is called **ergodic** if

$$P_t f \to \mu f = \mathbb{E}_\mu[f]$$

as $t \to +\infty$ for every $f \in L^2(\mu)$ (i.e. converges to the constant function $\mu f = \mu(f)$). That is, if we would like to measure $X_t \sim \rho_t$ with f, then far enough in time this measurement converges to measuring $X \sim \mu$ with f. Since this applies to all f (think $f = 1_A$), we can determine that $\rho_t \to \mu$ as $t \to +\infty$.

The following theorem determines whether a chain is ergodic, but note that we don't know anything about the *rate of convergence* to the stationary measure.

Theorem 4.5 ()

If Markov process $\{X_t\}$ with stationary measure μ and semigroup (P_t) is irreducible, then (P_t) is ergodic.

5 Martingales

Let us first start with the discrete-time martingale for simplicity. In introductory courses, a martingale might be defined as a stochastic process satisfying

$$X_n = \mathbb{E}[X_{n+1} \mid X_0, \dots, X_n]$$

for all n, which models a "fair game." They also may construct the random variables $\{X_n\}$ first and then define the filtration as the sequence of σ-algebras $\sigma(X_1, \dots, X_n)$. In here, we will construct the filtration $\{\mathcal{F}_n\}$ first and then define the random variables to be adapted to the filtration if X_n is $\mathcal{F}_n$-measurable for each $n \in \mathbb{N}$.

Definition 5.1 (Discrete-Time Martingale)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, let $\mathbb{F} = \{\mathcal{F}_n\}_{n\in\mathbb{N}}$ be a filtration (an increasing sequence of σ-algebras). A sequence $\{X_n\}$ is said to be **adapted** to $\{\mathcal{F}_n\}$ if X_n is $\mathcal{F}_n$-measurable for all n. If the stochastic process $\{X_n\}_{n\in\mathbb{N}}$ is a sequence with

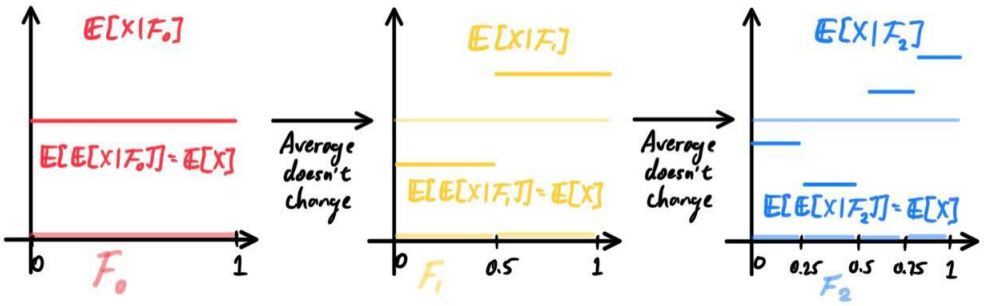
1. $\mathbb{E}[X_n] < \infty$ for all n,
2. X_n is adapted to $\mathcal{F}_n$,
3. $\mathbb{E}[X_{n+1} \mid X_1, \dots, X_n] = \mathbb{E}[X_{n+1} \mid \mathcal{F}_n] = X_n$ for all n,

then $\{X_n\}$ is a **martingale**. If $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \leq X_n$ or $\mathbb{E}[X_{n+1} \mid \mathcal{F}_n] \geq X_n$, the $\{X_n\}$ is said to be a **supermartingale** or **submartingale**, respectively.

A martingale just represents a sequence of random variables that get finer and finer as the σ-algebra increases. While they do get finer and finer, they do not change the "average" of the function. For example, consider the filtration generated by finer subsets of the unit interval $\Omega = (0, 1]$. We have

1. $\mathcal{F}_0 = \{\emptyset, \Omega\}$
2. $\mathcal{F}_1 = \sigma((0, 0.5], (0.5, 1])$
3. $\mathcal{F}_2 = \sigma((0, 0.25], (0.25, 0.5], (0.5, 0.75], (0.75, 1])$

Then, we would have



A supermartingale (and submartingale) just means that as we make the function finer and finer, its mean goes down (or up).

Martingales are used to model lots of random walk events. In the following three examples, let $\xi_1, \xi_2, \dots$ be iid, and let $S_n = S_0 + \xi_1 + \ldots + \xi_n.$, where S_0 is a constant. Let $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ for $n \geq 1$ and let $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

Example 5.1 (Linear Martingale)

Let $\mu = \mathbb{E}[\xi_i] = 0$. Then, $\{S_n\}$ is a martingale with respect to $\mathcal{F}_n$. We show the three requirements:

1. $\mathbb{E}[S_n] = \mathbb{E}[S_0] + \mathbb{E}[\xi_1] + \ldots + \mathbb{E}[\xi_n] = S_0 < \infty$.
2. By definition, we know that ξ_i is $\sigma(\xi)$-measurable for all $i \in [n]$, so ξ_i is $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$-measurable. Since the set of $\mathcal{F}_n$-measurable functions has a vector space structure, S_n is also $\mathcal{F}_n$-measurable.
3. We can simply solve

$$\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[S_n \mid \mathcal{F}_n] + \mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = X_n + \mathbb{E}[\xi_{n+1}] = X_n$$

where the first equality follows from linearity. For the second equality, note that S_n is $\mathcal{F}_n$-measurable from above, and so the best $\mathcal{F}_n$-measurable approximation of X is X itself (i.e. we have complete information). We know that ξ_{n+1} is independent of the ξ_i's, and so by definition their σ-algebras are independent. This implies that $\sigma(\xi_{n+1})$ and $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n)$ are independent, and so due to irrelevant information, $\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[\xi_{n+1}]$.

If $\mu \leq 0$ or $\mu \geq 0$, then the computation above shows that $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] \leq 0$ or $\mathbb{E}[S_{n+1} \mid \mathcal{F}_n] \geq 0$, making it a supermartingale or submartingale, respectively.

Given a supermartingale or submartingale, we can change it to be a martingale.

Example 5.2 ()

Given that $\mu = \mathbb{E}[\xi_i] \neq 0$, then $\{S_n - n\mu\}$is a martingale with respect to $\mathcal{F}_n$. We can see this because

$$\begin{aligned}
\mathbb{E}[S_{n+1} - (n+1)\mu \mid \mathcal{F}_n] &= \mathbb{E}[S_n - n\mu \mid \mathcal{F}_n] + \mathbb{E}[\xi_{n+1} - \mu \mid \mathcal{F}_n] \\
&= S_n - n\mu + \mathbb{E}[\xi_{n+1}] - \mu \\
&= S_n - n
\end{aligned}$$

Example 5.3 (Quadratic Martingale)

Say $\mu = \mathbb{E}[\xi_i] = 0$ and $\sigma^2 = \text{Var}(\xi_i) < \infty$. Then, $\{S_n^2 - n\sigma^2\}$ is a martingale.

$$\begin{aligned}
\mathbb{E}[S_{n+1}^2 - (n+1)\sigma^2 \mid \mathcal{F}_n] &= \mathbb{E}[(S_n + \xi_{n+1})^2 - (n_1)\sigma^2 \mid \mathcal{F}_n] \\
&= \mathbb{E}[S^2 - n\sigma^2 \mid \mathcal{F}_n] + \mathbb{E}[2S_n\xi_{n+1} + \xi_{n+1}^2 - \sigma^2 \mid \mathcal{F}_n] \\
&= \mathbb{E}[S^2 - n\sigma^2 \mid \mathcal{F}_n] + 2\mathbb{E}[S_n\xi_{n+1} \mid \mathcal{F}_n] + \mathbb{E}[\xi_{n+1}^2] - \sigma^2 \\
&= \mathbb{E}[S^2 - n\sigma^2 \mid \mathcal{F}_n]
\end{aligned}$$

where we have used the fact that due to independence of ξ_{n+1} with $\mathcal{F}_n$, we have $\mathbb{E}[S_n\xi_{n+1} \mid \mathcal{F}_n] = \mathbb{E}[S_n\mathbb{E}[\xi_{n+1} \mid \mathcal{F}_n]] = \mathbb{E}[S_n \cdot 0] = 0$.

This following result shows that martingales with bounded increments either converge or oscillate between $+\infty$ and $-\infty$.

Theorem 5.1 ()

Let $\{X_n\}_{n\in\mathbb{N}}$ be a martingale with $|X_{n+1} - X_n| \leq M < \infty$. Let

$$C = \{\lim_{n\to\infty} X_n \text{ exists and is finite}\}$$
$$D = \{\lim_{n\to\infty} \sup X_n = +\infty \text{ and } \lim_{n\to\infty} \inf X_n = -\infty\}$$

Then $\mathbb{P}(C \cup D) = 1$.

6 Concentration Inequalities

An informal statement of concentration of measure is the following: *If $X_1, \dots, X_n$ are independent random variables, then the random variable $f(X_1, \dots, X_n)$ is "close" to its mean $\mathbb{E}[f(X_1, \dots, X_n)]$ provided that the function $f(x_1, \dots, x_n)$ is not too "sensitive" to any of the coordinates x_i.* Intuitively, say that we have a bunch of independent random variables X_i and sample from them, to get some values x_i. Calculating $f(x_1, \dots, x_n)$, we have sampled from $f(X_1, \dots, X_n)$. Since f depends smoothly w.r.t. its arguments, to drastically change f, we must drastically change all the arguments. This is not likely, since all the X_i's are independent.

Most of our intuition about probability in low-dimensional spaces breaks down in high-dimensional ones (on the order of perhaps 10 or 20). We start off with two geometric examples in high-dimensional space.

Example 6.1 (Uniform Measure on Sphere)

Let μ_n be the uniform probability distribution on the n-sphere $\mathbf{S}^n \subset \mathbb{R}^{n+1}$. That is, let us consider any measurable set $A \subset \mathbb{S}^n$ such that $\mu_n(A) \geq 1/2$. Then, if we let $d(x, A)$ be the geodesic distance between $x \in \mathbb{S}^n$ and A , we define the expanded set

$$A_t = \{x \in \mathbb{S}^n \mid d(x, A) < t\}$$

and it turns out that

$$\mu_n(A_t) \geq 1 - e^{-(n-1)t^2/2}$$

which states that given *any* length $t > 0$, no matter how small, A_t almost covers the whole space. Then, for large enough n, μ_n is highly concentrated around the equator.

Note that the bounds decay *exponentially* (or of greater order).

Example 6.2 (Uniform Measure on Cube)

Example 6.3 (High Dimensional Gaussian)

Given iid $X_1, \dots, X_n \sim \mathcal{N}(0, \sigma^2)$, then let $\mathbf{X}$ be the random n-vector of these random variables. Then, the random variable

$$||\mathbf{X}|| = \sqrt{X_1^2 + \dots, X_n^2}$$

has a distribution that is very concentrated around the expectation

$$\mathbb{E}[||\mathbf{X}||] = \sqrt{\frac{n}{3}}$$

Naturally, this concentration phenomenon extends to random variables.

Example 6.4 ()

Let us have iid random variables X_i with $\mathbb{P}(X_i = 1) = 1/2$ and $\mathbb{P}(X_i = -1) = 1/2$. Then, let's define $S_n = \sum_{i=1}^n X_i$. The strong law of large numbers tell us that

$$\frac{S_n}{n} \xrightarrow{a.s.} 0$$

while the central limit theorem tells us that

$$\frac{S_n}{\sqrt{n}} \xrightarrow{d} \mathcal{N}(0, 1)$$

since $\mathbb{E}[X_i] = 0$ and $\text{Var}[X_i] = 1$. The CLT result shows us that the fluctuations (variance) of S_n of are order n. However, note that $|S_n|$ can take values as large as n, so the maximum value of S_n/n is of order 1. If we measure S_n using this scale, then $\frac{S_n}{n}$ is essentially 0. The actual bound looks like

$$\mathbb{P}\left(\frac{|S_n|}{n} \geq r\right) \leq 2e^{-nr^2/2}$$

Lemma 6.1 (Markov's Inequality)

Given any random variable X, we have

$$\mathbb{P}(X \geq \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

Lemma 6.2 (Chebyshev's Inequality)

Given X with finite variance and expectation, we have

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \alpha) \leq \frac{\text{Var}[X]}{\alpha^2}$$

An inequality that we will use often in proofs is Jensen's inequality.

Lemma 6.3 (Jensen's Inequality)

Given a convex function $g : \mathbb{R} \to \mathbb{R}$ and random variable X, we have

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]$$

Proof.

We will assume that f is differentiable for simplicity and let $\mathbb{E}[X] = \mu$. Define the linear function centered at μ to be $l(x) \coloneqq f(\mu) + f'(\mu)(x - \mu)$. Then, we know that $f(x) \geq l(x)$ for all x, so

$$\begin{aligned}
\mathbb{E}[f(X)] &\geq \mathbb{E}[l(X)] \\
&= \mathbb{E}[f(\mu) + f'(\mu)\,(X - \mu)] \\
&= \mathbb{E}[f(\mu)] + f'(\mu)(\mathbb{E}[X] - \mu) \\
&= \mathbb{E}[f(\mu)] \\
&= f(\mathbb{E}[X])
\end{aligned}$$

Definition 6.1 (Lipschitz Continuity)

A function $f : (X, d_X) \longrightarrow (Y, d_Y)$ is **Lipschitz continuous**, with Lipschitz constant A, if it satisfies

$$d_Y\big(f(\mathbf{x}), f(\mathbf{y})\big) \leq A\, d_X(\mathbf{x}, \mathbf{y})$$

for all $\mathbf{x}, \mathbf{y} \in X$.

6.1 Talagrand's Gaussian Inequality

Lemma 6.4 (Gaussian Integration by Parts Formula)

For Gaussian random variables $x, x_1, \dots, x_n$ and a function F of moderate growth at infinity, we have

$$\mathbb{E}\big[x\, F(x_1, \dots, x_n)\big] = \sum_{i=1}^{n} \mathbb{E}[x\, x_i]\, \mathbb{E}\left[\frac{\partial F}{\partial x_i}(x_1, \dots, x_n)\right]$$

Theorem 6.1 (Talagrand's Gaussian Inequality)

Consider a Lipschitz function $F : \mathbb{R}^N \longrightarrow \mathbb{R}$ (with Lipschitz constant A). Let $x_1, \dots, x_N \sim \mathcal{N}(0, 1)$ be iid, and let $\mathbf{x} = (x_1, \dots, x_N)$. Then, for each $t > 0$, we have

$$\mathbb{P}\big(|F(\mathbf{x}) - \mathbb{E}F(\mathbf{x})| \geq t\big) \leq 2 \exp\left(-\frac{t^2}{4A^2}\right)$$

Proof.

For this proof, we assume that F is not only Lipschitz, but C^2. This is the case in most applications of this theorem, and if it is not the case, then we can regularize F by convolving with a smooth function to solve the problem. We begin with a parameter s and consider the function $G : \mathbb{R}^{2N} \longrightarrow \mathbb{R}$ defined

$$G(z_1, \dots, z_{2N}) = \exp\Big(s\big[F(z_1, \dots, z_N) - F(z_{N+1}, \dots, z_{2N})\big]\Big)$$

For clarity, we will denote variables of F with x_i and variables of G with z_i. Let $u_1, \dots, u_{2N} \sim \mathcal{N}(0, 1)$ be iid, and let $v_1, \dots, v_n \sim \mathcal{N}(0, 1)$ be iid, with $v_{N+1}, \dots, v_{2N}$ copies of the first N. For shorthand, we can denote the collection as $\mathbf{u}$ and $\mathbf{v}$. Then, we have

$$\mathbb{E}[u_i u_j] - \mathbb{E}[v_i v_j] = 0$$

except when $j = i + M$ or $i = j + M$, in which case we have

$$\mathbb{E}[u_i u_j] - \mathbb{E}[v_i v_j] = 0 - 1 = -1$$

since $v_i v_j = X^2$, where $X \sim \mathcal{N}(0, 1) = \chi_1^2$, a Chi-Squared distribution with 1 degree of freedom. We consider the transformed random variable

$$\mathbf{f}(t) \coloneqq \sqrt{t}\, \mathbf{u} + \sqrt{1 - t}\, \mathbf{v} \sim \mathcal{N}(0, 1) \text{ for all } t$$

that is essentially some smooth path from $\mathbf{f}(0) = \mathbf{u}$ and $\mathbf{f}(1) = \mathbf{v}$. Note that given some $t \in [0, 1]$, $\mathbf{f}(t)$ is some random vector, $G(\mathbf{f}(t))$ is some random variable, and $\mathbb{E}[G(\mathbf{f}(t))]$ is some number. We can define the function $\phi : [0, 1] \longrightarrow \mathbb{R}$ as

$$\begin{aligned} \phi(t) = \mathbb{E}[G(\mathbf{f}(t))] &= \int_{\mathbb{R}} x\; p_{G(f(t))}(x)\, dx \\ &= \int_{\mathbb{R}^{2N}} G(y)\; p_{f(t)}(y)\, dy \end{aligned}$$

where p_X is the PDF of the distribution X. Take the derivative with respect to t to get the first line, and we can simplify using Gaussian integration by parts

$$\begin{aligned}
&\phi'(t)\mathbb{E}\Big[\sum_{i=1}^{2N}\frac{d}{dt}f_i(t)\,\frac{\partial G}{\partial z_i}\big(\mathbf{f}(t)\big)\Big] \\
&\quad=\sum_{i=1}^{2N}\mathbb{E}\Big[\frac{d}{dt}f_i(t)\,\frac{\partial G}{\partial z_i}\big(\mathbf{f}(t)\big)\Big] \\
&\quad=\sum_{i=1}^{2N}\sum_{j=1}^{2N}\mathbb{E}\Big[\Big(\frac{\partial}{\partial t}f_i(t)\Big)\,f_i(t)\Big]\,\mathbb{E}\Big[\frac{\partial^2 G}{\partial z_i\partial z_j}\mathbf{f}(t)\Big]
\end{aligned}$$

But we can simplify

$$\begin{aligned}
\mathbb{E}\Big[\Big(\frac{\partial}{\partial t}f_i(t)\Big)\,f_i(t)\Big]&=\mathbb{E}\Big[\Big(\frac{1}{2\sqrt{t}}u_i-\frac{1}{2\sqrt{1-t}}v_i\Big)\big(\sqrt{t}u_j-\sqrt{1-t}\,v_j\big)\Big] \\
&=\frac{1}{2}\big(\mathbb{E}[u_iu_j]-\mathbb{E}[v_iv_j]\big)=\begin{cases}-1 & \text{if } j=i+M, i=j+M\\ 0 & \text{else}\end{cases}
\end{aligned}$$

So, we can simplify the above to

$$\phi'(t)=-\mathbb{E}\Big[\sum_{i=1}^{N}\frac{\partial^2 G}{\partial z_i\,\partial z_{i+M}}\big(\mathbf{f}(t)\big)\Big]$$

and computing the second derivative using the chain rule gives

$$\begin{aligned}
\frac{\partial G}{\partial z_i}(\mathbf{z})&=\frac{\partial G}{\partial F}\frac{\partial F}{\partial x_i}(z_1,\dots,z_N)\\
&=s\,G(\mathbf{z})\,\frac{\partial F}{\partial x_i}(z_1,\dots,z_N)\\
\frac{\partial^2 G}{\partial z_i\partial z_{i+N}}(\mathbf{z})&=-s^2\,G(\mathbf{z})\,\frac{\partial F}{\partial x_i}(z_1,\dots,z_N)\,\frac{\partial F}{\partial x_i}(z_{N+1},\dots,z_{2N})
\end{aligned}$$

for all $\mathbf{z}$. So we have for all $t\in[0,1]$,

$$\begin{aligned}
\phi'(t)&=s^2\,\mathbb{E}\Big[\sum_{i=1}^{N}G(\mathbf{f}(t))\,\frac{\partial F}{\partial x_i}\big(f_1(t),\dots,f_N(t)\big)\,\frac{\partial F}{\partial x_i}\big(f_{N+1}(t),\dots,f_{2N}(t)\big)\Big]\\
&\le s^2\mathbb{E}\Big[G(\mathbf{f}(t))\sum_{i=1}^{N}\frac{\partial F}{\partial x_i}\big(f_1(t),\dots,f_N(t)\big)\,\frac{\partial F}{\partial x_i}\big(f_{N+1}(t),\dots,f_{2N}(t)\big)\Big]\\
&\le s^2\mathbb{E}\big[G(\mathbf{f}(t))\,A^2\big]\\
&\le s^2A^2\mathbb{E}[G(\mathbf{f}(t))]=s^2A^2\phi(t)
\end{aligned}$$

Solving the inequality for ϕ gives

$$\begin{aligned}
\phi'(t)/\phi(t)\le s^2A^2\implies&\int\phi'(t)/\phi(t)\,dt\le\int s^2A^2\,dt\\
\implies&\log\phi(t)\le s^2A^2t+C\\
\implies&\phi(t)\le e^{s^2A^2t}\le e^{s^2A^2}
\end{aligned}$$

Recalling that $\mathbf{f}(1)=\mathbf{u}$, we have

$$\mathbb{E}[\exp\{s(F(u_1,\dots,u_N)-F(u_{N+1},\dots,u_{2N}))\}]\le e^{s^2A^2}$$

and by independence of the u_i's, the LHS equals $\mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}]$ and by Jensen's inequality, we have $\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}] \geq e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}$. We can derive as follows:

$$\begin{aligned} e^{s^2A^2} &\geq \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-sF(u_{N+1},\dots,u_{2N})}] \\ &\geq \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]} \\ &= \mathbb{E}[e^{sF(u_1,\dots,u_N)}]\,\mathbb{E}[e^{-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}] \\ &= \mathbb{E}[e^{sF(u_1,\dots,u_N)-s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]}] \\ &= \mathbb{E}[\exp\big(sF(u_1,\dots,u_N) - s\mathbb{E}[F(u_{N+1},\dots,u_{2N})]\big)] \end{aligned}$$

and by Markov's inequality, we get for a random vector of standard Gaussian random variables $\mathbf{x}$

$$\begin{aligned} \mathbb{P}\big(F(\mathbf{x}) - \mathbb{E}[F(\mathbf{x})] \geq t\big) &= \mathbb{P}\big(e^{s(F(\mathbf{x})-\mathbb{E}[F(\mathbf{x})]} \geq e^{st}\big) \\ &\leq \frac{\mathbb{E}[e^{s(F(\mathbf{x})-\mathbb{E}[F(\mathbf{x})]}]}{e^{st}} \\ &\leq e^{s^2A^2-st} \\ &= e^{-t^2/4A^2} \text{ when } s = t/2A^2 \end{aligned}$$

7 Variance Bounds and Poincare Inequalities

Let us first describe this concentration phenomenon by investigating bounds on the variance

$$\mathrm{Var}[f(x_1,\dots,x_n)] \coloneqq \mathbb{E}\big[\big(f(x_1,\dots,x_n) - \mathbb{E}[f(x_1,\dots,x_n)]\big)^2\big]$$

We can first bound

$$\mathrm{Var}[f(X_1,\dots,X_n)] = \mathbb{E}\big[\big(f(X_1,\dots,X_n)\big)^2\big] - \mathbb{E}\big[f(X_1,\dots,X_n)\big]^2 \leq \mathbb{E}\big[\big(f(X_1,\dots,X_n)\big)^2\big]$$

and since adding a constant term to f doesn't affect the variance, we can utilize this to get our first variance bound.

Lemma 7.1 ()

Let $\mathbf{X}$ be a random variable or vector. Then,

$$\mathrm{Var}[f(\mathbf{X})] \leq \mathbb{E}\big[\big(f(\mathbf{X}) - \inf f\big)^2\big] \text{ and } \mathrm{Var}[f(\mathbf{X})] \leq \mathbb{E}\big[\big(\sup f - f(\mathbf{X})\big)^2\big]$$

and

$$\mathrm{Var}[f(\mathbf{X})] \leq \frac{1}{4}(\sup f - \inf f)^2$$

Proof.

Since $\mathrm{Var}[\mathbf{X}] = \mathbb{E}[\mathbf{X}^2] - \mathbb{E}[\mathbf{X}]^2$ from above, we have

$$\mathrm{Var}[f(\mathbf{X})] = \mathrm{Var}[f(\mathbf{X}) - a] = \mathbb{E}[(f(\mathbf{X}) - a)^2] - \mathbb{E}[f(\mathbf{X}) - a]^2 \leq \mathbb{E}[(f(\mathbf{X}) - a)^2]$$

By letting $a = \inf f$, we get the first inequality. By letting $a = (\sup f + \inf f)/2$ be the "middle" of f, we have $|f(\mathbf{X}) - a| \leq (\sup f - \inf f)/2 \implies [f(\mathbf{X}) - a]^2 \leq (\sup f - \inf f)^2/4$, and so

$$\mathrm{Var}[f(\mathbf{X})] \leq \mathbb{E}[(f(\mathbf{X}) - a)^2] \leq \frac{1}{4}(\sup f - \inf f)^2$$

which gives our third inequality. We can also see that

$$\text{Var}[f(\mathbf{X})] = \text{Var}[-f(\mathbf{X})] = \text{Var}[b - f(\mathbf{X})] \leq \mathbb{E}[(b - f(\mathbf{X}))^2]$$

to get our second.

This allows us to bound the random vector $f(\mathbf{X})$ if f itself is bounded, no matter what $\mathbf{X}$ is. But this generally turns out to be a very conservative bound, which is unsurprising since we assume so little about $\mathbf{X}$. For example, if we let $X_1, \dots, X_n$ be iid random variables taking values in $[-1, 1]$, and let $f(x_1, \dots, x_n) = \frac{1}{n}\sum_{i=1}^n x_i$. Then, f takes values in $[-1, 1]$, and by the previous lemma, we have

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}(1 - (-1))^2 = 1$$

which looks good, until we see that we can derive a better bound from direct computation (which becomes much better as n increases).

$$\text{Var}[f(X_1, \dots, X_n)] = \frac{1}{n^2}\sum_{i=1}^n \text{Var}[X_i] = \frac{1}{n}$$

However, this computation assumes independence of X_i's, which the previous lemma doesn't. This is the reason we're able to get a better bound, since if we took n copies of the same X, we would have

$$\text{Var}[f(X_1, \dots, X_n)] = \text{Var}[nX/n] = \text{Var}[X] = 1$$

Therefore, we will capitalize on the independence of these random variables in high dimensions to obtain better bounds. Now in the next result, we shall show that the variance of a high dimensional $f(X_1, \dots, X_n)$ can be bounded by the variances of each random variable. Those quantities, like the variance, that behave well in high dimensions is said to *tensorize*.

Consider independent random variables $X_1, \dots, X_n$ and a function $f : \mathbb{R}^n \longrightarrow \mathbb{R}$. If we fix values $x_1, \dots, x_n$, then we can define for all $k = 1, \dots, n$ the function $g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n) : \mathbb{R} \to \mathbb{R}$ as

$$g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(z) = f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

where

$$(g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n))'(z) = \frac{\partial}{\partial x_k} f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

and $g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)$ is a random variable of X_k. Then, we can define

$$\begin{aligned}
\text{Var}_k f(x_1, \dots, x_n) &= \text{Var}_{X_k}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1} \dots, x_n)] \\
&= \mathbb{E}_{X_k}\left[\left(f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n) - \mathbb{E}_{X_k}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n)]\right)^2\right] \\
&= \text{Var}[g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)] \\
&= \text{Var}_{X_k}[g(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)]
\end{aligned}$$

which takes the variance of f with respect to X_k, keeping all other variables fixed. However, this value will change for different $x_1, \dots, x_n$'s, and so we can loosen the restriction that they are fixed. We can take

$$g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(z) = f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n)$$

where $g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)$ is a random variable of $X_1, \dots, X_n$. Now if we calculate its partial variance, we get

$$\begin{aligned}
\text{Var}_k f(X_1, \dots, X_n) &= \text{Var}_{X_k}[f(X_1, \dots, X_k, \dots, X_n)] \\
&= \text{Var}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)] \\
&= \text{Var}_{X_k}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)]
\end{aligned}$$

which is now a random variable of all X_i's, $i \neq k$, that outputs the variance of f with respect to X_k. **But is it true that**

$$\mathbb{E}_{X_k}[f(X_1, \dots, X_n)] = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]?$$

Now, we can show a very useful property of variance: that the variance of some arbitrary function can be bounded by the expected sum of the partial variances.

Theorem 7.1 (Tensorization of Variance)

That is, $\operatorname{Var}_i f(\mathbf{x})$ is the variance of $f(X_1, \dots, X_n)$ w.r.t. the variable X_i only, the remaining variables kept fixed. Then, we have

$$\operatorname{Var}[f(X_1, \dots, X_n)] \leq \mathbb{E}\left[\sum_{i=1}^{n} \operatorname{Var}_i f(X_1, \dots, X_n)\right]$$

Proof.

We try to mimic the fact that the variance of the sum of independent random variables is the sum of the variances. At first sight, the general function $f(x_1, \dots, x_n)$ need not look anything like a sum, but we can expand it as a telescoping sum of random variables. We will prove this using the *martingale method*, which constructs this random variable $f(X_1, \dots, X_n)$ as a sum of finer and finer increments starting from the "coarse" constant function $\mathbb{E}[f(X_1, \dots, X_n)]$. We define the random variable

$$\Delta_k \coloneqq \mathbf{E}[f(X_1, \dots, X_n) \mid X_1, \dots X_k] - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

Then, we can express

$$f(X_1, \dots, X_n) - \mathbf{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^{n} \Delta_k$$

Note that $\mathbb{E}[\Delta_k \mid X_1, \dots, X_{k-1}] = 0$ (i.e. Δ_k's are martingale increments). In particular, even though the Δ_k's are not independent, if we have $l < k$, then

$$\begin{aligned}
\mathbb{E}[\Delta_k \Delta_l] &= \mathbb{E}[\mathbb{E}[\Delta_k \Delta_l \mid X_1, \dots, X_{k-1}]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_k \mid X_1, \dots X_{k-1}]\, \mathbb{E}[\Delta_l \mid X_1, \dots X_{k-1}]] \\
&= \mathbb{E}[\mathbb{E}[\Delta_k \mid X_1, \dots X_{k-1}]\, \Delta_l] \\
&= \mathbb{E}[0 \cdot \Delta_l] = 0
\end{aligned}$$

and so, the variance can be expanded into terms that vanish.

$$\begin{aligned}
\operatorname{Var}[f(X_1, \dots, X_n)] &= \mathbb{E}\big[\big(f(X_1, \dots, X_n) - \mathbf{E}[f(X_1, \dots, X_n)]\big)^2\big] \\
&= \mathbb{E}\left[\left(\sum_{k=1}^{n} \Delta_k\right)^2\right] = \sum_{k=1}^{n} \mathbb{E}[\Delta_k^2]
\end{aligned}$$

Now it remains to show that $\mathbb{E}[\Delta_k^2] \leq \mathbb{E}[\operatorname{Var}_k f(X_1, \dots, X_n)]$ for every k. Let us define

$$\tilde{\Delta}_k = f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]$$

to be the approximation of $f(X_1, \dots, X_n)$ "one step" before the final increment. Then, we have

$$\Delta_k = \mathbb{E}[\tilde{\Delta}_k \mid X_1, \dots, X_k]$$

and as X_k and $X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n$ are independent, we have

$$\operatorname{Var}_k f(X_1, \dots, X_n) = \mathbb{E}[\tilde{\Delta}_k^2 \mid X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n]$$

and therefore using Jensen's inequality we can prove

$$\mathbb{E}[\Delta_k^2] = \mathbb{E}[\mathbb{E}[\tilde{\Delta}_k \mid X_1, \dots, X_k]^2] \leq \mathbb{E}[\tilde{\Delta}_k^2] = \mathbb{E}[\operatorname{Var}_k f(X_1, \dots, X_n)]$$

What we want to eventually do is prove an inequality of the form where for any function $h : \mathbb{R} \to \mathbb{R}$ and some $X \sim \mu$,

$$\text{Var}_\mu[h] = \text{Var}[h(X)] \leq ||\mathcal{L}(h)||^2_{L^2(\mu)}$$

where $\mathcal{L}$ is an operator on h. This will allow us to bound

$$\text{Var}[g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)(X_k)] \leq ||\mathcal{L}(g_k(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n))||^2$$

for all $x_1, \dots, x_n$, simply by taking $h = g(x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n)$. Since this works for all $x_1, \dots, x_n$, we can claim that this inequality holds for all $X_1(\omega), \dots, X_n(\omega)$ for all $\omega \in \Omega$. That is, we can loosen the fixed values into random variables.

$$\begin{aligned}\text{Var}_k\, f(X_1, \dots, X_n) &= \text{Var}[g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n)(X_k)] \\ &\leq ||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}\end{aligned}$$

Note that all terms are random variables of $X_1, \dots, X_n$, and so the same inequality holds for their expectations over the entire joint measure.

$$\mathbb{E}[\text{Var}_k\, f(X_1, \dots, X_n)] \leq \mathbb{E}\big[||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}\big]$$

and so by tensorization (i.e. summing them up), we get

$$\text{Var}[f(X_1, \dots, X_n)] \leq \sum_{i=1}^n \mathbb{E}\big[\,\text{Var}_i\, f(X_1, \dots, X_n)\big] \leq \sum_{i=1}^n \mathbb{E}\big[||\mathcal{L}(g_k(X_1, \dots, X_{k-1}, X_{k+1}, \dots, X_n))||^2_{L^2(\mu)}\big]$$

Furthermore, this bound is sharp when f is linear. Let us demonstrate this by letting $f(x_1, \dots, x_n) = a_1x_1 + \dots + a_nx_n$. On the left hand side, we have

$$\text{Var}[f(X_1, \dots, X_n)] = \text{Var}\left[\sum_{i=1}^n a_iX_i\right] = \sum_{i=1}^n a_i^2\,\text{Var}[X_i]$$

and on the right hand side, each component divides up to

$$\begin{aligned}\text{Var}_i\, f(x_1, \dots, x_n) &= \text{Var}[f(x_1, \dots, X_i, \dots, x_n)] \\ &= \text{Var}[a_1x_1 + \dots + a_iX_i + \dots a_nx_n] \\ &= \text{Var}[a_iX_i] \\ &= a_i^2\,\text{Var}[X_i]\end{aligned}$$

Then? Note that since f is linear, the values of all $x_j, j \neq i$ have no effect on the variance of X_i, and so $\text{Var}_i\, f(X_1, \dots, X_n)$, which is originally a random variable of $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$, is really just the constant (random variable) $a_i^2\,\text{Var}[X_i]$. This is because no matter what values $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ are realized, these values will only contribute to a translation of the random variable $f(X_1, \dots, X_n)$, and hence will not affect the variance w.r.t. X_i. So, the right hand side also becomes

$$\mathbb{E}\left[\sum_{i=1}^n \text{Var}_i\, f(X_1, \dots, X_n)\right] = \mathbb{E}\left[\sum_{i=1}^n a_i^2\,\text{Var}[X_i]\right] = \sum_{i=1}^n a_i^2\,\text{Var}[X_i]$$

which is the same as the LHS.

We can view the tensorization of the variance in itself as an expression of the concentration phenomenon. $\text{Var}_i\, f(\mathbf{x})$ quantifies the sensitivity of the function $f(\mathbf{x})$ of the coordinate x_i in a distribution-dependent manner. If this sensitivity w.r.t. each coordinate ($\mathbb{E}[\text{Var}_i\, f(X_1, \dots, X_n)]$) is small, then $f(X_1, \dots, X_n)$ is close to its mean. However, it might not be so straightforward to compute $\text{Var}_i\, f$, since it depends on both the function f and on the distribution of X_i. So, we can try combining this with a suitable bound on the component-wise variance.

Let us define the quantities:

$$D_i f(\mathbf{x}) \coloneqq \sup_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n) - \inf_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

and

$$D_i^- f(\mathbf{x}) \coloneqq f(x_1, \dots, x_n) - \inf_z f(x_1, \dots, x_{i-1}, z, x_{i+1}, \dots, x_n)$$

which quantifies the sensitivity of the function f to the coordinate x_i in a distribution-independent manner. Now we can introduce the following bounds.

Corollary 7.1 ()

We have

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \big(D_i f(X_1, \dots, X_n)\big)^2\right]$$

Proof.

We start off with

$$\begin{aligned}\text{Var}_i f(X_1, \dots, X_n) &= \text{Var}[f(X_1, \dots, X_i, \dots, X_n)] \\ &\leq \frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\end{aligned}$$

Since these a random variables follow this inequality (for all $\omega \in \Omega$), we can attach an expectation on them to get

$$\mathbb{E}[\text{Var}_i f(X_1, \dots, X_n)] \leq \mathbb{E}\left[\frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\right]$$

and substituting in the previous theorem gives

$$\begin{aligned}\text{Var}[f(X_1, \dots, X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \text{Var}_i f(X_1, \dots, X_n)\right] \\ &= \sum_{i=1}^n \mathbb{E}\big[\text{Var}_i f(X_1, \dots, X_n)\big] \\ &\leq \sum_{i=1}^n \mathbb{E}\left[\frac{1}{4}\big(D_i f(X_1, \dots, X_n)\big)^2\right] \\ &= \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \big(D_i f(X_1, \dots, X_n)\big)^2\right]\end{aligned}$$

Example 7.1 (Random Matrices)

Exercise 7.1 (Banach-Valued Sums)

Let $X_1, X_2, \dots, X_N$ be independent random variables with values in a Banach space $(B, ||\cdot||_B)$. Suppose these random variables are bounded in the sense that $||X_i||_B \leq C$ a.s. for every i. Show that

$$\text{Var}\left(\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right|_B\right) \leq \frac{C^2}{n}$$

This is a simple vector-valued variant of the elementary fact that the variance of $\frac{1}{n}\sum_{k=1}^n X_k$ for

real-valued random variables X_k is of order $\frac{1}{n}$.

Solution 7.1

We can tensorize the variance to get

$$\begin{aligned}\operatorname{Var}_k \left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right|_B &= \operatorname{Var}\left|\left|\frac{1}{n}X_k\right|\right|_B = \frac{1}{n^2}\operatorname{Var}||X_k||_B \\ &\leq \frac{1}{n^2}\left(\frac{1}{4}(C-(-C))^2\right) = \frac{C^2}{n^2}\end{aligned}$$

and so letting $f(X_1,\dots,X_n) = \left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right|_B$, we get

$$\begin{aligned}\operatorname{Var}[f(X_1,\dots X_n)] &\leq \sum_{k=1}^n \mathbb{E}[\operatorname{Var}_k f(X_1,\dots,X_n)] \\ &\leq \sum_{k=1}^n \frac{C^2}{n^2} = \frac{C^2}{n}\end{aligned}$$

Exercise 7.2 (Rademacher Processes)

Let $\epsilon_1,\dots,\epsilon_n$ be independent symmetric Bernoulli random variables $\mathbb{P}(\epsilon_i = \pm 1) = \frac{1}{2}$ (also called Rademacher variables), let $T \subset \mathbb{R}^n$. The following identity is completely trivial:

$$\sup_{t\in T}\operatorname{Var}\left[\sum_{k=1}^n \epsilon_k t_k\right] = \sup_{t\in T}\sum_{k=1}^n t_k^2$$

Prove the following nontrivial fact:

$$\operatorname{Var}\left[\sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k\right] \leq 4\sup_{t\in T}\sum_{k=1}^n t_k^2$$

Solution 7.2

Let us consider a fixed $\boldsymbol{\epsilon} = (\epsilon_1,\dots,\epsilon_n)$ and index $i \in [n]$. Then, consider the random variable formed by taking the value $f(\epsilon_1,\dots,\epsilon_n)$ and loosening ϵ_i to be an random variable. That is,

$$\begin{aligned}\mathbb{P}\Big[f(\epsilon_1,\dots,\epsilon_n) = \sup_{t\in T}\{\epsilon_1 t_1 + \dots + 1t_i + \dots + \epsilon_n t_n\}\Big] &= \frac{1}{2} \\ \mathbb{P}\Big[f(\epsilon_1,\dots,\epsilon_n) = \sup_{t\in T}\{\epsilon_1 t_1 + \dots - 1t_i + \dots + \epsilon_n t_n\}\Big] &= \frac{1}{2}\end{aligned}$$

Then, we compute

$$D_i^- f(\epsilon_1,\dots,\epsilon_n) = \inf_{\epsilon_i\in\{-1,1\}}\sup_{t\in T}\sum_{k=1}^n \epsilon_k t_k$$

and we can estimate

$$\begin{aligned} D_i^- f(\boldsymbol{\epsilon}) &= f(\epsilon_1, \dots, \epsilon_n) - D_i f(\epsilon_1, \dots, \epsilon_n) \\ &= \sup_{t \in T} \sum_{k=1}^n \epsilon_k t_k - \inf_{\epsilon_i \in \{-1,1\}} \sup_{t \in T} \sum_{k=1}^n \epsilon_k t_k \\ &\leq \sup_{t \in T} 2|t_i| \end{aligned}$$

We can finally bound

$$\begin{aligned} \text{Var}[f(\epsilon_1, \dots, \epsilon_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \left(D_i^- f(\boldsymbol{\epsilon}) \right)^2 \right] \\ &\leq 4\mathbb{E}\left[\sum_{i=1}^n \sup_{t \in T} t_i^2 \right] \\ &= 4 \sup_{t \in T} \sum_{i=1}^n t_i^2 \end{aligned}$$

Exercise 7.3 (Bin Packing)

This is a classical application of bounded difference inequalities. Let $X_1, \dots, X_n$ i.i.d. random variables with values in $[0, 1]$. Each X_i represents the size of a package to be shipped. The shipping containers are bins of size 1 (so each bin can hold a set packages whose sizes sum to at most 1). Let $B_n = f(X_1, \dots, X_n)$ be the minimal number of bins needed to store the packages. Note that computing B_n is a hard combinatorial optimization problem, but we can bound its mean and variance by easy arguments.

1. Show that $\text{Var}[B_n] \leq n/4$
2. Show that $\mathbb{E}[B_n] \geq n\mathbb{E}[X_1]$

Thus the fluctuations $\sim \sqrt{n}$ of B_n are much smaller than its magnitude $\sim n$.

Solution 7.3

Listed.

1. Given fixed sizes $X_1, \dots, X_n$ and some $i \in [n]$, we can see that a property of f is that

$$f(X_1, \dots, X_{i-1}, 0, X_{i+1}, \dots, X_n) + 1 = f(X_1, \dots, X_{i-1}, 1, X_{i+1}, \dots, X_n)$$

since for an extra package with size 1, you would for sure need one more bin. So the maximum difference of f based on the x_i value is the constant random variable

$$\begin{aligned} D_i f(X_1, \dots, X_n) &= \sup_{z \in [0,1]} f(X_1, \dots, z, \dots, X_n) - \inf_{z \in [0,1]} f(X_1, \dots, z, \dots, X_n) \\ &= f(X_1, \dots, 1, \dots, X_n) - f(X_1, \dots, 0, \dots, X_n) = 1 \end{aligned}$$

and so by the bounded difference inequalities,

$$\begin{aligned} \text{Var}[B_n] = \text{Var}[f(X_1, \dots, X_n)] &\leq \frac{1}{4}\mathbb{E}\left[\sum_{i=1}^n \left(D_i f(X_1, \dots, X_n) \right)^2 \right] \\ &= \frac{1}{4} \sum_{i=1}^n \mathbb{E}\left[\left(D_i f(X_1, \dots, X_n) \right)^2 \right] \\ &\leq \frac{n}{4} \end{aligned}$$

2. Given the sizes $X_1, \dots, X_n$, B_n must satisfy

$$B_n = f(X_1, \dots, X_n) \geq X_1 + \dots + X_n$$

since the total volume of bins B_n must exceed the total volume $X_1 + \dots + X_n$ of packages. So,

$$\mathbb{E}[B_n] \geq \mathbb{E}\left[\sum_{k=1}^{n} X_k\right] = n\mathbb{E}[X_1]$$

Exercise 7.4 (Order Statistics and Spacings)

Let $X_1, \dots, X_n$ be independent random variables, and denote by $X_{(1)} \geq \dots \geq X_{(n)}$ their decreasing rearrangement ($X_{(1)} = \max_i X_i$, $X_{(n)} = \min_i X_i$, etc.). Show that

$$\text{Var}[X_{(k)}] \leq k\,\mathbb{E}[(X_{(k)} - X_{(k+1)})^2] \text{ for } 1 \leq k \leq n/2$$

and that

$$\text{Var}[X_{(k)}] \leq (n-k+1)\,\mathbb{E}[(X_{(k-1)} - X_{(k)})^2] \text{ for } n/2 < k \leq n$$

Exercise 7.5 (Convex Poincare Inequality)

Let $X_1, \dots, X_n$ be independent random variables taking values in $[a, b]$. The bounded difference inequalities estimate the variance $\text{Var}[f(X_1, \dots, X_n)]$ in terms of *discrete* derivatives $D_i f$ or $D_i^- f$ of the function f. The goal of this problem is to show that if the function f is convex, then one can obtain a similar bound in terms of the ordinary notion of derivative $\nabla_i f(x) = \partial f(x)/\partial x_i$ in $\mathbb{R}^n$.

1. Show that if $g : \mathbb{R} \longrightarrow \mathbb{R}$ is convex, then

$$g(y) - g(x) \geq g'(x)\,(y - x) \text{ for all } x, y \in \mathbb{R}$$

2. Show using part (a) and the bounded difference inequalities that if $f : \mathbb{R}^n \to \mathbb{R}$ is convex, then

$$\text{Var}[f(X_1, \dots, X_n)] \geq (b-a)^2 \mathbb{E}[||\nabla f(X_1, \dots, X_n)||^2]$$

3. Conclude that if f is convex and L-Lipschitz, i.e. $|f(x) - f(y)| \leq L||x - y||$ for all $x, y \in [a, b]^n$, then $\text{Var}[f(X_1, \dots, X_n)] \geq L^2(b-a)^2$.

Solution 7.4

Listed.

1. Assuming g is differentiable, let us choose any $x, y \in \mathbb{R}$ and define some $z = \lambda x + (1-\lambda)y$ in between. Then, pictorially, we would like to formally show that

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$

and take the limit as $z \to x$ to get $f'(x)$ on the LHS. By definition, we have

$$f(z) = f\big(\lambda x + (1-\lambda)y\big) \leq \lambda f(x) + (1-\lambda) f(y)$$

Subtracting $f(x)$ and then dividing by $1 - \lambda > 0$ on both sides gives

$$\frac{f(z) - f(x)}{1 - \lambda} \leq f(y) - f(x)$$

Note that $z - x = \lambda x + (1-\lambda y) - x = (1-\lambda)(y-x)$. So, dividing by $y - x > 0$ on both sides gives

$$\frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$

and taking the limit on the LHS gives

$$f'(x) = \lim_{z \to x} \frac{f(z) - f(x)}{z - x} \leq \frac{f(y) - f(x)}{y - x}$$

Since $y - x > 0$, we can multiply both on the same side to get

$$f(y) - f(x) \geq f'(x)\,(y - x)$$

If $y < x$, then the proof is the same, and the inequality sign ends up getting switched around twice, leading to the same conclusion.

2. Note that from the above result, we can multiply both sides by -1 to get that $g(x) - g(y) \leq g'(x)(x-y)$ for all $x, y \in \mathbb{R}$, and then swap the two variables to get $g(y) - g(x) \leq g'(y)(y-x)$. Let us consider fixed $x_1, \dots, x_n$ and some $i \in [n]$. Given $f : \mathbb{R}^n \to \mathbb{R}$, we define $f_i(\mathbf{x}) : \mathbb{R} \to \mathbb{R}$ by unfixing the ith variable. Then, given some $\alpha, \beta \in [a, b]$,

$$f_i(\mathbf{x})(\beta) - f_i(\mathbf{x})(\alpha) \leq g'(\beta)(\beta - \alpha)$$

or equivalently,

$$f(x_1, \dots, \beta, \dots, x_n) - f(x_1, \dots, \alpha, \dots, x_n) \leq \frac{\partial f}{\partial x_i}(x_1, \dots, \beta, \dots, x_n)\,(\beta - \alpha)$$

Now let $z^* \in [a, b]$ be the value s.t.

$$z^* = \arg\min_{z \in [a,b]} f(x_1, \dots, z, \dots, x_n)$$

Then,

$$D_i^- f(\mathbf{x}) = f(x_1, \dots, x_i, \dots x_n) - f(x_1, \dots, z^*, \dots, x_n) \leq \frac{\partial f}{\partial x_i}(x_1, \dots, x_i, \dots, x_n)\,(x_i - z^*)$$

and so

$$\left(D_i^- f(\mathbf{X})\right)^2 \leq \nabla_i f(\mathbf{x})^2\,(x_i - z^*)^2 \leq \nabla_i f(\mathbf{x})^2\,(b-a)^2$$

which gives from the bounded difference inequality

$$\begin{aligned}
\text{Var}[f(X_1, \dots, X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \left(D_i^- f(X_1, \dots, X_n)\right)^2\right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n \nabla_i f(\mathbf{x})^2\,(b-a)^2\right] \\
&= (b-a)^2 \mathbb{E}\left[\left|\left|\nabla f(\mathbf{X})\right|\right|^2\right]
\end{aligned}$$

3. If f is L-lipschitz, then $||\nabla f(\mathbf{X})|| \leq L$, and so

$$\text{Var}[f(X_1, \dots, X_n)] \leq (b-a)^2 L^2$$

7.1 Markov Semigroups

Definition 7.1 (Markov Process)

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $(S, \mathcal{S})$ be a measurable space. A homogeneous Markov process $\{X_t\}_{t\geq 0}$ is a stochastic process that satisfies the **Markov property**: for every bounded measurable function f and $s, t \geq 0$, there exists a bounded measurable function $P_s f$ satisfying

$$\mathbb{E}[f(X_{t+s}) \mid \{X_r\}_{r\leq t}] = (P_s f)(X_t) = \mathbb{E}[f(X_{t+s}) \mid X_t]$$

Definition 7.2 (Stationary Measure)

A probability measure μ is called **stationary** or **invariant** if

$$\mathbb{E}_\mu[f] = \mathbb{E}_\mu[P_t f] \text{ i.e. } \int_S f \, d\mu = \int_S P_t f d\mu$$

for all $t \geq 0$ and bounded measurable f. By abusing notation, this is conventionally written

$$\mu(f) = \mu(P_t f)$$

To interpret this notion, suppose that $X_0 \sim \mu$. Then,

$$\mathbb{E}[f(X_t)] = \mathbb{E}[\mathbb{E}[f(X_t) \mid X_0]] = \mathbb{E}[P_t f(X_0)] = \mathbb{E}_\mu[P_t f]$$

and if μ is stationary, then we have $\mathbb{E}[f(X_t)] = \mathbb{E}_\mu[f]$. If $f = 1_A$ for some measurable $A \subset S$, then $\mathbb{E}[1_A(X_t)] = \mathbb{P}(X_t \in A)$, and

$$\mathbb{P}(X_t \in A) = \mathbb{E}_\mu[1_A] = \int_S 1_A \, d\mu = \int_A d\mu = \mu(A) = \mathbb{P}(X_0 \in A)$$

which means that the probability that for all $A \in \mathcal{S}$ and all $t \geq 0$, the probability of X_t realizing in A is equivalent to the initial probability of X_0 realizing in A. This means that the process remains distributed according to the stationary measure $X_t \sim \mu$ for every time t. In summary, stationary measures describe the equilibrium or steady-state behavior of the Markov process.

From now, given the state space $(S, \mathcal{S})$ we can put a measure μ on it to get a measure space $(S, \mathcal{S}, \mu)$. The Banach space of all μ-measurable functions $f : (S, \mathcal{S}, \mu) \to (\mathbb{R}, \mathcal{R})$ (i.e. for every Borel $B \in \mathcal{R}$, $f^{-1}(B) \in \mathcal{S}$) will be denoted $L^p(\mu)$, equipped with the norm

$$||f||_{L^p(\mu)} \coloneqq \mathbb{E}_\mu[f^p]^{1/p} = \left(\int_S |f|^p \, d\mu \right)^{1/p}$$

If $p = 2$, then we can define the inner product

$$\langle f, g \rangle_\mu \coloneqq \mathbb{E}_\mu[fg] = \int_S fg \, d\mu$$

Lemma 7.2 ()

Let μ be a stationary measure. Then, the following hold for all $p \geq 1$, $t, s \geq 1$, $\alpha, \beta \in \mathbb{R}$, and bounded measurable functions f, g.

1. Contraction:

$$||P_t f||_{L^p(\mu)} \leq ||f||_{L^p(\mu)} = \mathbb{E}_\mu[f^p]^{1/p}$$

2. Linearity:
$$P_t(\alpha f + \beta g) = \alpha P_t f + \beta P_t g$$
3. Semigroup Property:
$$P_{t+s} f = P_t P_s f$$
4. Conservativeness:
$$P_t 1 = 1$$

Lemma 7.3 ()

Let μ be a stationary measure. Then, $t \mapsto \text{Var}_\mu[P_t f]$ is a decreasing function of time for every function $f \in L^2(\mu)$.

Proof.

Note that

$$\begin{aligned}\text{Var}_\mu[P_t f] &= ||P_t f - \mu f||^2_{L^2(\mu)} = ||P_t(f - \mu f)||^2_{L^2(\mu)} = ||P_{t-s}P_s(f - \mu f)||^2_{L^2(\mu)} \\ &\leq ||P_s(f - \mu f)||^2_{L^2(\mu)} = ||P_s f - \mu f||^2_{L^2(\mu)} = \text{Var}_\mu(P_s f)\end{aligned}$$

We now define the analogous operator to the transition rate matrix in discrete time chains with a finite state space.

Definition 7.3 (Generator)

The generator $\mathscr{L}$ is defined as
$$\mathscr{L} f \coloneqq \lim_{t \downarrow 0} \frac{P_t f - f}{t}$$
for every $f \in L^2(\mu)$ for which the above limit exists in $L^2(\mu)$. The set of f for which $\mathscr{L} f$ is defined is called the domain $\text{Dom}(\mathscr{L})$ of the generator, and $\mathscr{L}$ defines a linear operator from $\text{Dom}(\mathscr{L}) \subset L^2(\mu)$ to $L^2(\mu)$.

We have defined the generator $\mathscr{L}$ from the Markov semigroup $\{P_t\}_{t \geq 0}$. Now, let's try to define the semigroup in terms of the generator $\mathscr{L}$. Given that we have some map $\mathscr{L}$), can we define some semigroup $\{P_t\}$ satisfying the definition? To do this, we must solve the differential equation:

$$\frac{d}{dt} P_t = \lim_{\delta \downarrow 0} \frac{P_{t+\delta} - P_t}{\delta} = \lim_{\delta \downarrow 0} \frac{P_t P_\delta - P_t}{\delta} = P_t \lim_{\delta \downarrow 0} \frac{P_\delta - I}{\delta} = P_t \mathscr{L}$$

For function P_t to satisfy this differential equation, we have the solution

$$P_t = e^{t\mathscr{L}}$$

which also implies that $\mathscr{L}$ and P_t must commute.

Definition 7.4 (Reversibility)

The Markov semigroup $\{P_t\}_{t \geq 0}$ with stationary measure μ is called **reversible** if

$$\langle f, P_t g \rangle_\mu = \langle P_t f, g \rangle_\mu$$

for every $f, g \in L^2(\mu)$. Equivalently, we can say that P_t is self-adjoint on $L^2(\mu)$, or since $P_t = e^{t\mathscr{L}}$, we have $\mathscr{L}$ is self-adjoint.

Definition 7.5 (Ergodicity)

The Markov semigroup $\{P_t\}_{t\geq 0}$ with stationary measure μ if called **ergodic** if

$$P_t f \to \mu f$$

in $L^2(\mu)$ as $t \to +\infty$ for every $f \in L^2(\mu)$. Note that $\mu f = \mu(f)$ is the constant function in $L^2(\mu)$.

Exercise 7.6 (Elementary Identities)

Let P_t be a Markov semigroup with generator $\mathscr{L}$ and stationary measure μ. Prove the following elementary facts.

1. Show that $\mu(\mathscr{L} f) = 0$ for every $f \in L^2(\mu)$
2. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $P_t\phi(f) \geq \phi(P_t f)$ when $f, \phi(f) \in L^2(\mu)$
3. If $\phi : \mathbb{R} \to \mathbb{R}$ is convex, then $\mathscr{L}\phi(f) \geq \phi'(f)\mathscr{L} f$ when $f, \phi(f) \in L^2(\mu)$
4. Let $f \in L^2(\mu)$. Show that the following process is a martingale.

$$M_t^f := f(X_t) - \int_0^t \mathscr{L} f(X_s)\, ds$$

Solution 7.5

Listed.

1. This is simply a property of the generator. Not worrying about interchanging limits and integrals, we have

$$\begin{aligned}\mu(\mathscr{L} f) = \mathbb{E}_\mu[\mathscr{L} f] &= \int_S \lim_{t\downarrow 0} \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \int_S \frac{P_t f - P_0 f}{t}\, d\mu \\ &= \lim_{t\downarrow 0} \frac{1}{t}\big(\mathbb{E}_\mu[P_t f] - \mathbb{E}_\mu[f]\big) = \lim_{t\downarrow 0} \frac{1}{t} \cdot 0 = 0\end{aligned}$$

2. By Jensen's inequality,

$$\begin{aligned}P_s\phi(f) &= \mathbb{E}[\phi(f)(X_{t+s}) \mid X_t] \\ &\geq \phi\Big(\mathbb{E}[f(X_{t+s} \mid X_t]\big) = \phi(P_s f)\end{aligned}$$

7.2 Poincare Inequalities

Recall that a Poincare inequality for μ is, informally, of the form

$$\text{variance}(f) \leq \mathbb{E}_\mu[||\text{gradient}(f)||^2]$$

At first sight, such an inequality has nothing to do with Markov processes. However, the validity of a Poincare inequality for μ turns out to be related to the rate of convergence of an ergodic Markov process for which μ is the stationary distribution. That is, a measure μ satisfies a Poincare inequality for a certain notion of gradient if and only if an ergodic Markov semigroup associated to this gradient converges exponentially fast to μ.

Definition 7.6 (Dirichlet Form)

Given a Markov process with generator $\mathscr{L}$ and stationary measure μ, the corresponding Dirichlet form is defined as

$$\mathcal{E}(f,g) \coloneqq -\langle f, \mathscr{L} g\rangle_\mu$$

Theorem 7.2 (Poincare Inequality)

Let P_t be a reversible ergodic Markov semigroup with stationary measure μ. The following are equivalent given $c \geq 0$.

1. $\text{Var}_\mu(f) \leq c\mathcal{E}(f,f)$ for all f (Poincare Inequality)
2. $||P_t f - \mu f||_{L^2(\mu)} \leq e^{-t/c}||f - \mu f||_{L^2(\mu)}$
3. $\mathcal{E}(P_t f, P_t f) \leq e^{-2t/c}\mathcal{E}(f,f)$ for all f, t
4. For every f there exists $\kappa(f)$ s.t. $||P_t f - \mu f||_{L^2(\mu)} \leq \kappa(f)e^{-t/c}$
5. For every f there exists $\kappa(f)$ s.t. $\mathcal{E}(P_t f, P_t f) \leq \kappa(f)e^{-2t/c}$

We should view properties 2 through 5 as different notions of exponential convergence of the Markov semigroup P_t to the stationary measure μ. Properties 2 and 4 directly measure the rate of convergence of $P_t f$ to μf in $L^2(\mu)$, while properties 3 and 5 measure the rate of convergence of the "gradient" (now depicted as $\mathcal{E}$) of $P_t f$ to 0.

7.2.1 The Gaussian Poincare Inequality

Definition 7.7 (Ornstein-Uhlenbeck Process)

Given standard Brownian motion $(W_t)_{t\geq 0}$, the **Ornstein-Uhlenbeck process** is defined as

$$X_t = e^{-t}X_0 + e^{-t}W_{e^{2t}-1}$$

Lemma 7.4 (Gaussian Integration by Parts)

If $\xi \sim \mathcal{N}(0,1)$, then

$$\mathbb{E}[\xi f(\xi)] = \mathbb{E}[f'(\xi)]$$

Proof.

Assuming that f is smooth with compact support, we have by integration by parts

$$\begin{aligned}
\mathbb{E}[f'(\xi)] &= \int_{-\infty}^{\infty} f'(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \\
&= \frac{e^{-x^2/2}}{\sqrt{2\pi}} f(x)\Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} f(x)\frac{d}{dx}\left(\frac{e^{-x^2/2}}{\sqrt{2\pi}}\right)dx \\
&= -\int_{-\infty}^{\infty} -xf(x)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx \\
&= \int_{-\infty}^{\infty} \big(xf(x)\big)\frac{e^{-x^2/2}}{\sqrt{2\pi}}\,dx = \mathbb{E}[\xi f(\xi)]
\end{aligned}$$

Theorem 7.3 ()

The Ornstein-Uhlenbeck Process $(X_t)_{t\geq 0}$

1. is a Markov process with semigroup
$$P_t f(x) = \mathbb{E}\big[f(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\big] \text{ with } \xi \sim \mathcal{N}(0,1)$$
2. admits $\mu = \mathcal{N}(0,1)$ as its stationary measure
3. is ergodic
4. has generator and Dirichlet form given by
$$\mathscr{L} f(x) = -xf'(x) + f''(x), \qquad \mathcal{E}(f,g) = \langle f', g' \rangle_\mu$$
5. is reversible

Proof.

Let $s \geq t$.

1. By definition of X_t, we have $X_t = e^{-t}X_0 + e^{-t}W_{e^{2t-1}}$ and
$$X_s = e^{-s}X_0 + e^{-s}W_{e^{2s}-1} \implies X_0 = (X_s - e^{-s}W_{e^{2s}-1})e^s$$
Substituting in the equation for X_s gives
$$\begin{aligned} X_t &= e^{-(t-s)}X_s + e^{-t}(W_{e^{2t}-1} - W_{e^{2s}-1}) \\ &= e^{-(t-s)}X_s + \sqrt{1-e^{-2(t-s)}}\xi \end{aligned}$$
where $\xi = (W_{e^{2t}-1} - W_{e^{2s}-1})/\sqrt{e^{2t}-e^{2s}} \sim N(0,1)$ is independent of $\{X_r\}_{r\leq s}$. Therefore, we can write
$$\mathbb{E}[f(X_t) \mid \{X_r\}_{r\leq s}] = P_{t-s}f(X_s) = \mathbb{E}\big[f\big(e^{-(t-s)}X_s + \sqrt{1-e^{-2(t-s)}}\xi\big)\big]$$
which proves the Markov property and gives the semigroup.
2. We can clearly see that if $X_t \sim N(0,1)$, then $X_{t+s} = e^{-s}X_t + \sqrt{1-e^{-2s}}\xi$ is a sum of Gaussians, one with variance e^{-2s} and the other with variance $1-e^{-2s}$, and so their sum has variance 1.
3. We will take for granted that this is ergodic.
4. To compute the generator, we use the chain rule (and not worry about whether we take the derivative within the expectation integral) and then use Gaussian integration by parts to get
$$\begin{aligned} \frac{d}{dt}P_t f(x) &= \mathbb{E}\left[f'(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\left(\frac{e^{-2t}}{\sqrt{1-e^{-2t}}}\xi - e^{-t}x\right)\right] \\ &= \mathbb{E}\big[e^{-t}xf'(e^{-t}x + \sqrt{1-e^{-2t}}\xi) + e^{-2t}f''(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\big] \end{aligned}$$
and therefore have
$$\frac{d}{dt}P_t f(x) = \left(-x\frac{d}{dx} + \frac{d^2}{dx^2}\right)P_t f(x)$$
The Dirichlet form can be simplified using the Gaussian integration by parts as
$$\begin{aligned} \mathcal{E}(f,g) &= -\langle f, \mathscr{L} g\rangle_\mu \\ &= \mathbb{E}[f(\xi)\big(xg'(\xi) - g''(\xi)\big)] \\ &= \mathbb{E}[\xi f(\xi)g'(\xi)] - \mathbb{E}[f(\xi)g''(\xi)] \\ &= \mathbb{E}[f'(\xi)g'(\xi) + f(\xi)g''(\xi)] - \mathbb{E}[f(\xi)g''(\xi)] \\ &= \mathbb{E}[f'(\xi)g'(\xi)] \end{aligned}$$
5. Since $\mathcal{E}(f,g) = \mathbb{E}[f'(\xi)g'(\xi)]$, it is symmetric and so $\mathscr{L}$ is self-adjoint.

From the previous theorem part 4, we can see that

$$\mathcal{E}(f,f) = \langle f', f' \rangle_\mu = ||f'||^2_{L^2(\mu)} = \mathbb{E}_\mu[f'^2]$$

which means that the Dirichlet form of an Ornstein-Uhlenbeck process is precisely the expected square gradient of function f! Therefore, with the Poincare inequality, we can bound the variance of f with the Dirichlet form, which is the expected square gradient of f.

Theorem 7.4 ()

Let $\mu = \mathcal{N}(0,1)$. Then,

$$\text{Var}_\mu[f] \leq ||f'||^2_{L^2(\mu)}$$

Proof.

We have from the properties of the Ornstein-Uhlenbeck process that

$$\begin{aligned}
\frac{d}{dx} P_t f(x) &= \frac{d}{dx} \mathbb{E}[f(e^{-t}x + \sqrt{1-e^{-2t}}\xi)] \\
&= \mathbb{E}\left[\frac{d}{dx} f(e^{-t}x + \sqrt{1-e^{-2t}}\xi) \right] \\
&= \mathbb{E}[f'(e^{-t}x + \sqrt{1-e^{-2t}}\xi)\ e^{-t}] \\
&= e^{-t}\mathbb{E}[f'(e^{-t}x + \sqrt{1-e^{-2t}}\xi)] \\
&= e^{-t} P_t f'(x)
\end{aligned}$$

Thus

$$\mathcal{E}(P_t f, P_t f) = ||(P_t f)'||^2_{L^2(\mu)} = e^{-2t}||P_t f'||^2_{L^2(\mu)} \leq e^{-2t}||f'||^2_{L^2(\mu)} = e^{-2t}\mathcal{E}(f,f)$$

where the inequality follows from contraction.

By tensorization, we can prove the following.

Corollary 7.2 (Gaussian Poincare Inequality)

Let $X_1, \dots, X_n \sim N(0,1)$ be iid. Then,

$$\text{Var}[f(X_1, \dots, X_n)] \leq \mathbb{E}[||\nabla f(X_1, \dots, X_n)||^2]$$

Proof.

Computation.

$$\begin{aligned}
\text{Var}[f(X_1, \dots, X_n)] &\leq \mathbb{E}\left[\sum_{i=1}^n \text{Var}_i f(X_1, \dots, X_n) \right] \\
&\leq \mathbb{E}\left[\sum_{i=1}^n \left|\left| \frac{d}{dx_i} f(X_1, \dots, X_n) \right|\right|^2 \right] \\
&= \mathbb{E}[||\nabla f(X_1, \dots, X_n)||^2]
\end{aligned}$$

So what have we done so far? If we have some distribution μ and want to prove an inequality that bounds $\text{Var}_\mu[f]$, then we should choose some (reversible ergodic) Markov process that has a stationary distribution μ. We can identify its semigroup, generator, and ultimately its Dirichlet form $\mathcal{E}(f,g)$, which will allow us to invoke the Poincare inequality to bound

$$\text{Var}_\mu[f] \leq c\mathcal{E}(f,f)$$

and since $\mu = N(0,1)$, we have shown above using both the properties of the generator of the Ornstein-Uhlenbeck process and Gaussian integration by parts that this Dirichlet form is precisely the norm of f'. This is clear since the Dirichlet form $\langle f, \mathscr{L} g \rangle_\mu$ only depends on $\mathscr{L}$ and μ. However, the Dirichlet form does not have to be this form.

1. If μ is some other distribution, we would not be able to reduce $\mathcal{E}(f,f)$ to the norm of its derivative, and so it make take on a different form.
2. If we choose a different Markov process, even with the same stationary measure $\mu = N(0,1)$, the generator may be different and so will the Dirichlet form.

Exercise 7.7 (Carre du Champ)

We have interpreted the Dirichlet form $\mathcal{E}(f,f)$ as a general notion of "expected square gradient" that arises in the study of Poincare inequalities. There is an analogous quantity $\Gamma(f,f)$ that plays the role of "square gradient" in this setting (without the expectation). In good probabilistic tradition, it is universally known by its French name carre du champ (literally, "square of the field"). The carre du champ is defined as

$$\Gamma(f,g) \coloneqq \frac{1}{2}\big[\mathscr{L}(fg) - f\mathscr{L}g - g\mathscr{L}f\big]$$

in terms of the generator $\mathscr{L}$ of a Markov process with stationary measure μ.

1. Show that $\mathcal{E}(f,f) = \int \Gamma(f,f)\, d\mu$ and that $\mathcal{E}(f,g) = \int \Gamma(f,g)\, d\mu$ if the Markov process is in addition reversible.
2. Show that $\Gamma(f,f) \geq 0$ so it can indeed by interpreted as a square.
3. Prove the Cauchy-Schwartz inequality $\Gamma(f,g)^2 \leq \Gamma(f,f)\,\Gamma(g,g)$
4. Compute the carre du champ of the Ornstein-Uhlenbeck process and confirm that it should indeed be interpreted as the appropriate notion of "square gradient."

Solution 7.6

Listed.

1. By stationarity, we have

$$\mu(\mathscr{L}f) = \int_S \mathscr{L}f\, d\mu = 0$$

for all $f \in L^2(\mu)$, which reduces the first term below to 0. So, we can reduce the carre du champ to

$$\begin{aligned}\int_S \Gamma(f,f)\, d\mu &= \frac{1}{2}\bigg(\int_S \mathscr{L}(f^2)\, d\mu - 2\int_S f\mathscr{L}f\, d\mu\bigg) \\ &= -\int_S f\mathscr{L}f\, d\mu = -\langle f, \mathscr{L}f\rangle_\mu = \mathcal{E}(f,f)\end{aligned}$$

Furthermore, assuming that P_t is reversible, we have

$$\mathcal{E}(f,g) = -\langle f, \mathscr{L}g\rangle_\mu = -\langle \mathscr{L}f, g\rangle_\mu = -\langle g, \mathscr{L}f\rangle_\mu = \mathcal{E}(g,f)$$

and so

$$\begin{aligned}\int \Gamma(f,g)\, d\mu &= \frac{1}{2}\bigg(\int \mathscr{L}(fg)\, d\mu - \int f\mathscr{L}g\, d\mu - \int g\mathscr{L}f\, d\mu\bigg) \\ &= \frac{1}{2}\big(-\langle f, \mathscr{L}g\rangle_\mu - \langle g, \mathscr{L}f\rangle_\mu\big) \\ &= -\langle f, \mathscr{L}g\rangle_\mu = \mathcal{E}(f,g)\end{aligned}$$

2. Since $\Gamma(f,f) = \frac{1}{2}\big(\mathscr{L}(f^2) - 2f\mathscr{L}f\big)$, the problem now reduces to proving that $\mathscr{L}(f^2) \geq 2f\mathscr{L}f$. By Jensen's inequality, we have $P_t(f^2) \geq (P_tf)^2$, and so

$$\begin{aligned}\mathscr{L}(f^2) &= \lim_{t\downarrow 0} \frac{P_t(f^2) - f^2}{t} \geq \lim_{t\downarrow 0} \frac{(P_tf)^2 - f^2}{t} \\ &= \frac{d}{dt}(P_tf)^2\bigg|_{t=0} = \left(2(P_tf)\cdot \frac{d}{dt}(P_tf)\right)\bigg|_{t=0} = 2f\mathscr{L}f\end{aligned}$$

3. We know that $\Gamma(f+tg, f+tg) \geq 0$ from above, and so if we expand out, we get

$$\begin{aligned}\Gamma(f+tg,f+tg) &= \frac{1}{2}\Big[\mathscr{L}\big((f+tg)^2\big) - 2(f+tg)\mathscr{L}(f+tg)\Big] \\ &= \Gamma(g,g)t^2 + 2\Gamma(f,g)t + \Gamma(f,f) \geq 0\end{aligned}$$

for all t. Since this quadratic is nonnegative, its discriminant must be ≤ 0, and so

$$\Delta = \big(2\Gamma(f,g)\big)^2 - 2\Gamma(g,g)\Gamma(f,f) \leq 0 \implies \Gamma(f,g)^2 \leq \Gamma(f,f)\Gamma(g,g)$$

4. The generator of the Ornstein-Uhlenbeck process is $\mathscr{L}f(x) = -xf'(x) + f''(x)$. Therefore,

$$\begin{aligned}\Gamma(f,g)(x) &= \frac{1}{2}\big[\mathscr{L}(fg)(x) - f(x)\mathscr{L}g(x) - g(x)\mathscr{L}f(x)\big] \\ &= \frac{1}{2}\Big[\big(-x(fg)'(x) + (fg)''(x)\big) - f(x)\big(-xg'(x) + g''(x)\big) - g(x)\big(-xf'(x) + f''(x)\big)\Big]\end{aligned}$$

which simplifies down to $f'(x)g'(x)$, and so $\Gamma(f,f) = [f'(x)]^2$ can be interpreted as the square gradient of f.

7.3 Variance Identities and Exponential Ergodicity

Now, let us develop some intuition on the connection between Markov semigroups, $\mathrm{Var}_\mu[f]$ and the Dirichlet form $\mathcal{E}(f,f)$.

Lemma 7.5 ()

The following identity holds.

$$\frac{d}{dt}\mathrm{Var}_\mu[P_tf] = -2\mathcal{E}(P_tf, P_tf)$$

Proof.

By stationarity, $\mu(P_tf) = \mu(f)$, and so

$$\begin{aligned}\frac{d}{dt}\mathrm{Var}_\mu[P_tf] &= \frac{d}{dt}\big\{\mu((P_tf)^2) - \mu(P_tf)^2\big\} \\ &= \frac{d}{dt}\big\{\mu((P_tf)^2) - \mu(f)^2\big\} = \frac{d}{dt}\mu((P_tf)^2) \\ &= \frac{d}{dt}\int_S (P_tf)^2\,d\mu = \int_S \frac{d}{dt}(P_tf)^2\,d\mu = 2\int_S (P_tf)\,\frac{d}{dt}P_tf\,d\mu \\ &= 2\mathbb{E}_\mu[P_tf, \mathscr{L}(P_tf)] = 2\langle P_tf, \mathscr{L}P_tf\rangle_\mu = -2\mathcal{E}(P_tf, P_tf)\end{aligned}$$

Theorem 7.5 ()

$\mathcal{E}(f, f) \geq 0$ for every f.

Proof.

We know that $t \mapsto \text{Var}_\mu[P_t f]$ is a decreasing function of t (by contraction of P_t), so

$$\frac{d}{dt} \text{Var}_\mu[P_t f] = -2\mathcal{E}(P_t f, P_t f) \leq 0$$

Theorem 7.6 ()

Suppose that the Markov semigroup is ergodic. Then, we have for every f

$$\text{Var}_\mu[f] = 2 \int_0^\infty \mathcal{E}(P_t f, P_t f)\, dt$$

8 Subgaussian Concentration and log-Sobolev Inequalities

8.1 Subgaussian Variables and Chernoff Bounds

We should first consider how one might go about proving that a random variable satisfies a Gaussian tail bound. Most tail bounds in probability theory are proved using some form of Markov's inequality.

Lemma 8.1 (Markov's Inequality)

Given a nonnegative random variable X, we have

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}$$

which means that the probability that $X > \alpha$ goes down at least as fast as $1/\alpha$.

Markov's inequality is very conservative but very general, too. If we make further assumptions about the random variable X, we can often make stronger bounds. Chebyshev's inequality assumes a (possibly negative) random variable with finite variance and states that the probability will go down as $1/x^2$.

Theorem 8.1 (Chebyshev Inequality)

Given (possibly negative) random variable X, if $\mathbb{E}[X] = \mu < +\infty$ and $\text{Var}(X) = \sigma^2 < +\infty$, then for all $\alpha > 0$,

$$\mathbb{P}\big(|X - \mu| > k\sigma\big) \leq \frac{1}{k^2} \iff \mathbb{P}(|X - \mu| > \alpha) \leq \frac{\text{Var}[X]}{\alpha^2}$$

That is, the probability that X takes a value further than k standard deviations away from μ goes down by $1/k^2$. Therefore, if σ is small, then this bound will be small since there is more concentration in the mean.

Proof.

We apply Markov's inequality to the non-negative random variable $|X - \mu|$.

$$\mathbb{P}(|X-\mu| > \alpha) = \mathbb{P}(|X-\mu|^2 > \alpha^2) \leq \frac{\mathbb{E}(|X-\mu|^2)}{\alpha^2} = \frac{\text{Var}[X]}{\alpha^2}$$

since the numerator on the RHS is the definition of variance.

Using higher powers, we can obtain better and better bounds, but not exponential ones. To obtain these Gaussian tail bounds, we must use more sophisticated methods.

Lemma 8.2 (Chernoff Bound)

Define the log-moment generating function ψ of a random variable X and its Legendre dual ψ^* as

$$\psi_X(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}[X] \qquad \psi_X^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi_X(\lambda)\}$$

Then, the following is known as the **Chernoff bound**.

$$\mathbb{P}[X - \mathbb{E}[X] \geq t] \leq e^{-\psi_X^*(t)}$$

for all $t \geq 0$. We can lower bound it too with

$$\mathbb{P}[X - \mathbb{E}[X] \leq -t] \leq e^{-\psi_X^*(t)}$$

and union bounding them gives

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\psi_X^*(t)}$$

Proof.

We take some $\lambda \geq 0$ and given that the map $x \mapsto e^{\lambda x}$ is nondecreasing, we can exponentiate and then use Markov's inequality:

$$\mathbb{P}[X - \mathbb{E}[X] \geq t] = \mathbb{P}[e^{\lambda(X-\mathbb{E}[X])} \geq e^{\lambda t}] \leq e^{-\lambda t}\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = e^{-(\lambda t - \psi_X(\lambda))} \leq e^{-\psi_X^*(t)}$$

as the left hand does not depend on the choice of λ, we have the additional flexibility of tuning λ to get potentially better bounds. We can also use Chernoff bound on the random variable $-X$ to bound

$$\begin{aligned}
\mathbb{P}(X - \mathbb{E}[X] \leq -t) &= \mathbb{P}(-X - \mathbb{E}[-X] \geq t) \\
&= \mathbb{P}(e^{\lambda(-X+\mathbb{E}[X])} \geq e^{\lambda t}] \\
&\leq e^{-\lambda t}\mathbb{E}[e^{\lambda(-X+\mathbb{E}[X])}] \\
&= e^{-(\lambda t - \psi_{-X}(\lambda))} \leq e^{-\psi_{-X}^*(t)}
\end{aligned}$$

There seems to be a minor problem in the fact that $-\psi_X^*$ and $-\psi_{-X}^*$ are different, and so provide different bounds for the upper and lower tail. But note that $\psi_X(\lambda) = \psi_{-X}(-\lambda)$, and so their maximum will coincide and $\psi_X^*(t) = \psi_{-X}^*(t)$, allowing us to get the union bound.

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t] \leq 2e^{-\psi^*(t)}$$

To observe how the Chernoff bound can give rise to Gaussian tail bounds, let us first consider the case of an actual Gaussian random variable.

Example 8.1 ()

Let $X \sim N(\mu, \sigma^2)$. Then, $\mathbb{E}[e^{\lambda(X-\mathbb{E}[X])}] = e^{\lambda^2\sigma^2/2}$, so

$$\psi(\lambda) = \frac{\lambda^2\sigma^2}{2}, \quad \psi^*(t) = \sup_{\lambda \geq 0} \left\{\lambda t - \frac{\lambda^2\sigma^2}{2}\right\} = \frac{t^2}{2\sigma^2}$$

and by the Chernoff bound, we have $\mathbb{P}(X - \mathbb{E}[X] \geq t] \leq e^{-t^2/2\sigma^2}$.

Note that in order to get the tail bound, the fact that X is Gaussian was not actually important. It would suffice to assume that the log-MGF is bouded from above by a Gaussian.

Definition 8.1 (Subgaussian Random Variables)

A random variable is called σ^2-**subgaussian** if its log-MGF satisfies

$$\psi(\lambda) \leq \frac{\lambda^2\sigma^2}{2}$$

for all $\lambda \in \mathbb{R}$. The constant σ^2 is called the **variance proxy**.

Remember that if $\psi(\lambda)$ is the log-MGF of a random variable X, then $\psi(-\lambda)$ is the log-MGF of the random variable $-X$. For a σ^2-subgaussian random variable X, we can therefore apply the Chernoff bound to both the upper and lower tails and union bound to obtain

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq 2e^{-t/2\sigma^2}$$

We have only worked with Gaussians, which are trivially subgaussian. A nontrivial results is that every bounded random variable is subgaussian.

Lemma 8.3 (Hoeffding's Lemma)

Let $a \leq X \leq b$ a.s. for some $a, b \in \mathbb{R}$. Then,

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

That is, X is $(b-a)^2/4$-subgaussian.

Proof.

We assume without loss of generality that $\mathbb{E}[X] = 0$. Then, we have $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}]$, and we can compute

$$\psi'(\lambda) = \frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}, \quad \psi''(\lambda) = \frac{\mathbb{E}[X^2e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \left(\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]}\right)^2$$

and thus

$$\psi''(\lambda) = \int_\Omega X^2 \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P} - \left(\int_\Omega X \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P}\right)^2$$

can be interpreted as the variance of the random variable X under the twisted probability measure $d\mathbb{Q} = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]}\, d\mathbb{P}$. But $a \leq X \leq b$, so we can bound the variance by its infimum and suprememum

$\psi''(\lambda) = \text{Var}_{\mathbb{Q}}[X] \leq (b-a)^2/4$, and the fundamental theorem of calculus yields

$$\psi(\lambda) = \int_0^\lambda \int_0^\mu \psi''(\rho)\, d\rho\, d\mu \leq \frac{\lambda^2(b-a)^2}{8}$$

using $\psi(0) = 0$ and $\psi'(0)$.

Exercise 8.1 (Subgaussian Variables)

There are several different notions of random variables with a Gaussian tail that are all essentialy equivalent up to constants. The aim of this problem is to obtain some insight into these notions.

1. Show that if X is σ^2-subgaussian, then $\text{Var}[X] \leq \sigma^2$.
2. Show that for any increasing and differentiable function Φ,

$$\mathbb{E}[\Phi(|X|)] = \Phi(0) + \int_0^\infty \Phi'(t)\, \mathbb{P}(|X| \geq t)\, dt$$

In the following, we will assume for simplicity that $\mathbb{E}[X] = 0$. We now prove that the following three properties are equivalent for suitable constants σ, b, c: (1) X is σ^2-subgaussian; (2) $\mathbb{P}(|X| \geq t) \leq 2e^{-bt^2}$; and (3) $\mathbb{E}[e^{cX^2}] \leq 2$.

3. Show that if X is σ^2-subgaussian , then $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2\sigma^2}$
4. Show that if $\mathbb{P}(|X| \geq t) \leq 2e^{-t^2/2\sigma^2}$, then $\mathbb{E}[e^{X^2/6\sigma^2}] \leq 2$.
5. Show that if $\mathbb{E}[e^{X^2/6\sigma^2}] \leq 2$, then X is $18\sigma^2$-subgaussian.

In addition, the subgaussian property of X is equivalent to the fact that the moments of X scale as is the case for the Gaussian distribution.

6. Show that if X is σ^2-subgaussian, then $\mathbb{E}[X^{2q}] \leq (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$.
7. Show that if $\mathbb{E}[X^{2q}] \leq (4\sigma^2)^q q!$ for all $q \in \mathbb{N}$, then $\mathbb{E}[e^{X^2/8\sigma^2}] \leq 2$.

Solution 8.1

Listed.

1. We can expand out

$$\begin{aligned}\mathbb{E}[e^{\lambda(X-\mathbb{E}X}] &= \mathbb{E}\left[1 + \lambda(X - \mathbb{E}X) + \frac{\lambda^2}{2}(X-\mathbb{E}X)^2 + \dots\right] \\ &= 1 + \frac{\lambda^2}{2}\text{Var}[X] + o(\lambda^2) \\ &\leq e^{\lambda^2\sigma^2/2} = 1 + \frac{\lambda^2\sigma^2}{2} + o(\lambda^2)\end{aligned}$$

   which is true for all λ. Setting $\lambda = 0$, we get $\text{Var}[X] \leq \sigma^2$.
2. Unfinished.
3. Since X is σ^2 subgaussian, its log-MGF satisfies $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] \leq \frac{\lambda^2\sigma^2}{2} \implies -\psi(\lambda) \geq -\frac{\lambda^2\sigma^2}{2}$. Then, its Legendre dual is

$$\psi^*(t) = \sup_{\lambda \geq 0}\{\lambda t - \psi(\lambda)\} \geq \sup_{\lambda \geq 0}\{\lambda t - \frac{\lambda^2\sigma^2}{2}\} = \frac{t^2}{2\sigma^2}$$

   where we optimize the quadratic w.r.t. λ. Therefore, $-\psi^*(t) \leq -\frac{t^2}{2\sigma^2} \implies \mathbb{P}(X \geq t) \leq e^{-\psi^*(t)} \leq e^{-t^2/2\sigma^2}$.

4. By using the identity above with $\Phi(t) = e^{t^2/6\sigma^2}$, we have

$$\begin{aligned}
\mathbb{E}[e^{X^2/6\sigma^2}] &= \mathbb{E}[e^{|X|^2/6\sigma^2}] \\
&= e^{0^2/6\sigma^2} + \int_0^\infty e^{t^2/6\sigma^2} \frac{t}{3\sigma^2} \mathbb{P}(|X| \geq t)\, dt \\
&\leq 1 + \frac{1}{3t^2} \int_0^\infty t e^{t^2/6\sigma^2}\, 2e^{-t^2/2\sigma^2}\, dt \\
&= 1 + \frac{2}{3\sigma^2} \int_0^\infty t e^{-\frac{1}{3}\frac{t^2}{\sigma^2}}\, dt \\
&= 1 - \frac{1}{\sigma^2} \int_0^\infty \Big(-\frac{2}{3\sigma} t \Big) e^{-\frac{t^2}{3\sigma^2}}\, dt \\
&= 1 - e^{-\frac{t^2}{3\sigma^2}} \Big|_0^\infty \\
&= 1 - (0 - 1) = 2
\end{aligned}$$

5. Unfinished.
6. We know $X^{2q} = |X|^{2q}$ for all $q \in \mathbb{N}$. By setting $\Phi(t) = t^{2q}$ from the identity above, we can get

$$\mathbb{E}[|X|^{2q}] = 0^{2q} + \int_0^\infty (2q) t^{2q-1} \mathbb{P}(|X| \geq t)\, dt$$

and from (3), we get the first line, where we can just keep doing integration by parts:

$$\begin{aligned}
\mathbb{E}[|X|^{2q}] &\leq \int_0^\infty (2q) t^{2q-1} e^{-t^2/2\sigma^2}\, dt \\
&= 2(4q\sigma^2) \int_0^\infty (2q-2) t^{2q-3} e^{-t^2/2\sigma^2}\, dt \\
&= 2(4q\sigma^2)(4(q-1)\sigma^2) \int_0^\infty (2q-4) t^{2q-5} e^{-t^2/2\sigma^2}\, dt \\
&= \dots \\
&= 2(4q\sigma^2) \dots (4 \cdot 2\sigma^2) \int_0^\infty 2t e^{-t^2/2\sigma^2}\, dt \\
&= \prod_{k=1}^{q} (4k\sigma^2) = (4\sigma^2)^q q!
\end{aligned}$$

7. We can expand and from the inequality above, we get

$$\begin{aligned}
\mathbb{E}[e^{X^2/8\sigma^2}] = \mathbb{E}\left[1 + \frac{X^2}{8\sigma^2} + \frac{1}{2}\left(\frac{X^2}{8\sigma^2}\right)^2 + \dots \right] \\
&= 1 + \sum_{q=1}^{\infty} \frac{1}{(8\sigma^2)^q q!} \mathbb{E}[X^{2q}] \\
&\leq 1 + \sum_{q=1}^{\infty} \frac{1}{(8\sigma^2)^q q!} (4\sigma^2)^q q! \\
&= 1 + \sum_{q=1}^{\infty} \frac{1}{2^q} = 2
\end{aligned}$$

Exercise 8.2 (Tightness of Hoeffding's Lemma)

Show that the bound on Hoeffding's lemma is the best possible by consider $\mathbb{P}(X=a)=\mathbb{P}(X=b)=\frac{1}{2}$.

Solution 8.2

From computing the expectation

$$\mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] = e^{\lambda(a-\frac{a+b}{2})}\mathbb{P}(X=a) + e^{\lambda(b-\frac{a+b}{2})}\mathbb{P}(X=b) = \frac{1}{2}e^{\lambda\frac{a-b}{2}} + \frac{1}{2}e^{\lambda\frac{b-a}{2}}$$

we know that this is always less than $\lambda^2(b-a)^2/8$ for all λ. But setting $\lambda=0$ satisfies equality.

8.2 The Martingale Method

In this section, we will use the martingale method to derive useful results. Recall that in order to derive some property (like tensorization of variance) of $f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)]$, we can expand it as a telescoping sum of martingale differences

$$f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)] = \sum_{k=1}^{n}\Delta_k$$

where

$$\Delta_k = \mathbb{E}[f(X_1,\dots,X_n)\mid X_1,\dots,X_k] - \mathbb{E}[f(X_1,\dots,X_n)\mid X_1,\dots,X_{k-1}]$$

and then deriving bounds on each difference. Note that these are martingale differences because given the filtration $\mathbb{F}=\{\mathcal{F}_k=\sigma(X_1,\dots,X_k)\}$, the stochastic process

$$Y_k = \sum_{i=1}^{k}\Delta_i = \mathbb{E}[f(X_1,\dots,X_n)\mid X_1,\dots,X_k] - \mathbb{E}[f(X_1,\dots,X_n)]$$

is a martingale.

Lemma 8.4 (Azuma)

Let $\mathbb{F}=\{\mathcal{F}_k\}_{k\le n}$ be any filtration, and $\Delta_1,\dots,\Delta_n$ be random variables that satisfy the following properties for $k=1,\dots,n$.

1. Martingale Difference Property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\Delta_k\mid\mathcal{F}_{k-1}]=0$
2. Conditional Subgaussian Property: $\mathbb{E}[e^{\lambda\Delta_k}\mid\mathcal{F}_{k-1}]\le e^{\lambda^2\sigma_k^2/2}$ a.s.

Then, the sum $\sum_{k=1}^n\Delta_k$ is subgaussian with variance proxy $\sum_{k=1}^n\sigma_k^2$.

Proof.

For any $1\le k\le n$, we can compute

$$\mathbb{E}[e^{\lambda\sum_{i=1}^{k}\Delta_i}] = \mathbb{E}[e^{\lambda\sum_{i=1}^{k-1}\Delta_i}\mathbb{E}[e^{\lambda\Delta_k}\mid\mathcal{F}_{k-1}]] \le e^{\lambda^2\sigma_k^2/2}\,\mathbb{E}[e^{\lambda\sum_{i=1}^{k-1}\Delta_i}]$$

and by induction, this proof is finished. Note that $\mathbb{E}[e^{\lambda\Delta_k}\mid\mathcal{F}_{k-1}]\le e^{\lambda^2\sigma_k^2/2}$ can only hold if $\mathbb{E}[\Delta_k\mid\mathcal{F}_{k-1}]=0$.

What this lemma basically says is that if we decompose a random variable into martingale differences, and each martingale difference is conditionally subgaussian, then their sum is also subgaussian. Now, if we just assume that each of these martingale differences are bounded, then we can use Hoeffding's lemma on each of them to make them subgaussian, and then use Azuma's lemma to show that their sum is subgaussian. This is exactly what we do here.

Theorem 8.2 (Azuma-Hoeffding Inequality)

Let $\mathbb{F} = \{\mathcal{F}_k\}_{k \leq n}$ be any filtration, and let Δ_k, A_k, B_k satisfy the following properties for $k = 1, \dots, n$.

1. Martingale Difference Property: Δ_k is $\mathcal{F}_k$-measurable and $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = 0$
2. Predictable bounds: A_k, B_k are $\mathcal{F}_{k-1}$-measurable and $A_k \leq \Delta_k \leq B_k$ a.s.

Then, $\sum_{k=1}^n \Delta_k$ is subgaussian with variance proxy $\frac{1}{4}\sum_{k=1}^n ||B_k - A_k||_\infty^2$. In particular, we obtain for every $t \geq 0$ the tail bound

$$\mathbb{P}\left(\sum_{k=1}^n \Delta_k \geq t \right) \leq \exp\left(- \frac{2t^2}{\sum_{k=1}^n ||B_k - A_k||_\infty^2} \right)$$

The Azuma-Hoeffding's inequality is often applied in the following setting. Let $X_1, \dots, X_n$ be independent random variables s.t. $a \leq X_i \leq b$ for all i (we can interpret a and b as simply constant random variables). Then, let $\Delta_k = (X_k - \mathbb{E}[X_k])/n$ be martingale differences, which we can show that Δ_k is clearly $\mathcal{F}_k$-measurable and that by independence of X_i's, $\mathbb{E}[\Delta_k \mid \mathcal{F}_{k-1}] = \mathbb{E}[\Delta_k] = 0$. Therefore, we can show that its sum satisfies

$$\mathbb{P}\left(\frac{1}{n} \sum_{k=1}^n \{X_k - \mathbb{E}[X_k]\} \geq t \right) \leq e^{-2nt^2/(b-a)^2}$$

which is consistent with the central limit theorem.

Now we can return to the case of functions $f(X_1, \dots, X_n)$ of independent random variables. Recall that the discrete derivative is defined

$$D_k f(x) = \sup_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n) - \inf_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$

Theorem 8.3 (McDiarmid)

For $X_1, \dots, X_n$ independent, $f(X_1, \dots, X_n)$ is subgaussian with variance proxy $\frac{1}{4}\sum_{k=1}^n ||D_k f||^2$. That is,

$$\mathbb{P}\big[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t\big] \leq \exp\left(- \frac{2t^2}{\sum_{k=1}^n ||D_k f||_\infty^2} \right)$$

Proof.

We use the martingale method again to write

$$f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] = \sum_{k=1}^n \Delta_k$$

where

$$\Delta_k = \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] - \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

What we want to do is set some upper and lower bound on $\mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k]$, which will set bounds on Δ_k. We can do this by bounding f by the infimum and supremum w.r.t. each element, getting

$$\begin{aligned} &\mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_k] \\ &\quad \leq \mathbb{E}[f(X_1, \dots, X_n) \mid X_1, \dots, X_k] \\ &\quad\quad \leq \mathbb{E}[\sup_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_k] \end{aligned}$$

but by independence of X_k's, we have

$$\mathbb{E}[\inf_z f(X_1, \dots, z, \dots, X_n) \mid X_1, \dots, X_k] = \mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

So, setting

$$A_k = \mathbb{E}[\inf_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$
$$B_k = \mathbb{E}[\sup_z f(X_1, \dots, X_{k-1}, z, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_n) \mid X_1, \dots, X_{k-1}]$$

we have $A_k \leq \Delta_k \leq B_k$ for all k, and by Azuma-Hoeffding's inequality along with the fact that $||B_k - A_k|| \leq ||D_k f||_\infty$, we get

$$\mathbb{P}[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n)] \geq t] \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n ||B_k - A_k||_\infty^2}\right) \leq \exp\left(-\frac{2t^2}{\sum_{k=1}^n ||D_k f||_\infty^2}\right)$$

We should treat McDiarmid's inequality as a subgaussian form of the bounded difference inequality

$$\text{Var}[f(X_1, \dots, X_n)] \leq \frac{1}{4}\mathbb{E}\left[\sum_{k=1}^n \big(D_k f(X_1, \dots, X_n)\big)^2\right]$$

The bounded difference inequality says that the variance is controlled by the expectation of the square gradient of the function f. In contrast, McDiarmid's inequality asserts the stronger subgaussian inequality, but under the stronger condition that the variance proxy is controlled by a uniform upper bound on the square gradient rather than its expectation. This will be a recurring theme:

1. the expectation of the square gradient controls the variance
2. a uniform bound on the square gradient controls the subgaussian property

Note that McDiarmid's theorem is not satisfactory. The appropriate notion of a square gradient in both inequalities is the random variable $\sum_{k=1}^n |D_k f|^2$. To control the variance, we want to take its expectation $\mathbb{E}[\sum_{k=1}^n |D_k f|^2]$, and to control the upper bound of the square gradient, we simply want to take its supremum $||\sum_{k=1}^n |D_k f|^2||_\infty$. However, McDiarmid's inequality only yields control in terms of the larger quantity $\sum_{k=1}^n ||D_k f||_\infty^2$ (by triangle inequality), which gets worse in higher dimensions. Rather than taking the supremum of square gradient, we just take the supremum of each (squared) component and add them up, which may be much greater than the actual upper bound. Therefore, the martingale method is far too crude to capture this idea, and we will need new techniques for more refined bounds.

Exercise 8.3 (Bin Packing)

For the Bin packing problem previoulsly, show that the variance bound $\text{Var}[B_n] \leq n/4$ can be strengthened to a Gaussian tail bound

$$\mathbb{P}(|B_n - \mathbb{E}B_n| \geq t) \leq 2e^{-2t^2/n}$$

Solution 8.3

We can see that

$$D_k f(X_1, \dots, X_n) = f(X_1, \dots, X_{k-1}, 1, X_{k+1}, \dots, X_n) - f(X_1, \dots, X_{k-1}, 1, X_{k+1}, \dots, X_n) = 1$$

and by McDiarmid's inequality, we are done.

Exercise 8.4 (Rademacher Processes)

Exercise 8.5 (Sums in Hilbert Space)

Let $X_1, \dots, X_n$ be independent random variables with zero mean that map to a Hilbert space, and suppose that $||X_k|| \leq C$ a.s. for every k.

1. Show that for all $t \geq 0$,
$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right| \geq \mathbb{E}\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right| + t\right] \leq e^{-nt^2/2C^2}$$
2. Show that
$$\mathbb{E}\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right| \leq Cn^{-1/2}$$
3. Conclude that for all $t \geq Cn^{-1/2}$,
$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right| \geq t\right] \leq e^{-nt^2/8C^2}$$
4. Finally, argue that for all $t \geq 0$,
$$\mathbb{P}\left[\left|\left|\frac{1}{n}\sum_{k=1}^n X_k\right|\right| \geq t\right] \leq e^{-nt^2/8C^2}$$

8.3 The Entropy Method

In order to develop more sophisticated concentration inequalities, let us introduce another term that is used to measure the deviation of a random variable.

Definition 8.2 (Entropy)

The **entropy** of a nonnegative random variable Z is defined
$$\text{Ent}[Z] \coloneqq \mathbb{E}[Z \log Z] - \mathbb{E}[Z] \log \mathbb{E}[Z]$$

Lemma 8.5 (Herbst)

Suppose that random variable X satisfies
$$\text{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}] \text{ for all } \lambda \geq 0$$
Then, X is σ^2-subgaussian. That is,
$$\psi(\lambda) \coloneqq \log \mathbb{E}[e^{\lambda(X - \mathbb{E}[X])}] \leq \frac{\lambda^2 \sigma^2}{2} \text{ for all } \lambda \geq 0$$

Proof.

As $\psi(\lambda) = \log \mathbb{E}[e^{\lambda X}] - \lambda \mathbb{E}[X]$, we have
$$\frac{d}{d\lambda}\frac{\psi(\lambda)}{\lambda} = \frac{1}{\lambda}\frac{\mathbb{E}[Xe^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} - \frac{1}{\lambda^2}\log \mathbb{E}[e^{\lambda X}] = \frac{1}{\lambda^2}\frac{\text{Ent}[e^{\lambda X}]}{\mathbb{E}[e^{\lambda X}]} \leq \frac{\sigma^2}{2}$$

where the last inequality yields from the assumption. By the fundamental theorem of calculus, we have

$$\frac{\psi(\lambda)}{\lambda} = \lim_{\lambda \downarrow 0} \frac{\psi(\lambda)}{\lambda} + \int_0^\lambda \frac{1}{t^2} \frac{\text{Ent}[e^{tX}]}{\mathbb{E}[e^{tX}]}\, dt \leq \frac{\lambda \sigma^2}{2} \implies \psi(\lambda) \leq \frac{\lambda^2 \sigma^2}{2}$$

Exercise 8.6 ()

It turns out that the converse is true up to a constant: If X is $\frac{\sigma^2}{4}$-subgaussian, then

$$\text{Ent}[e^{\lambda X}] \leq \frac{\lambda^2 \sigma^2}{2} \mathbb{E}[e^{\lambda X}]$$

Solution 8.4

We know that by Jensen's inequality and concavity of the logarithm,

$$\log \mathbb{E}[e^{\lambda(X - \mathbb{E}X)}] \geq \mathbb{E}[\lambda(X - \mathbb{E}X)] = 0 \implies \mathbb{E}[e^{\lambda(X-\mathbb{E}X)}] \geq 1$$

Furthermore, note that given $Z = e^{\lambda X} / \mathbb{E}[e^{\lambda X}]$, we have

$$\begin{aligned}
\mathbb{E}[Z \log Z] &= \mathbb{E}\left[\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} \log \left(\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X}]} \right) \right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]} \mathbb{E}\left[e^{\lambda X} \left(\log e^{\lambda X} - \log \mathbb{E}[e^{\lambda X}] \right) \right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]} \mathbb{E}\left[e^{\lambda X} \lambda X - e^{\lambda X} \log \mathbb{E}[e^{\lambda X}] \right] \\
&= \frac{1}{\mathbb{E}[e^{\lambda X}]} \left(\mathbb{E}[e^{\lambda X} \lambda X] - \mathbb{E}[e^{\lambda X}] \log \mathbb{E}[e^{\lambda X}] \right) \\
&= \frac{\text{Ent}[e^{\lambda X}]}{\mathbb{E}[e^{[\lambda X}]}
\end{aligned}$$

Since this theorem assumes a bound on $\text{Ent}[e^{\lambda X}]$ rather than $\text{Ent}[X]$, we will mainly be working with the entropy of exponentials of a random variable.

It turns out that entropy behaves very similarly to variance and extends nicely into the subgaussian setting. Just like variance, we define the partial entropy of function $f(x_1, \dots, x_n)$ as

$$\text{Ent}_k f(x_1, \dots, x_n) \coloneqq \text{Ent}[f(x_1, \dots, x_{k-1}, X_k, x_{k+1}, \dots, x_n)]$$

That is, $\text{Ent}[f(X_1, \dots, X_n)]$ is the entropy of $f(X_1, \dots, X_n)$ with respect to the variable X_k only, the remaining variables kept fixed.

Theorem 8.4 (Tensorization of Entropy)

Given that $X_1, \dots, X_n$ are independent,

$$\text{Ent}[f(X_1, \dots, X_n)] \leq \mathbb{E}\left[\sum_{k=1}^n \text{Ent}_k f(X_1, \dots, X_n) \right]$$

Recall that the basic method for deriving Poincare inequalities is that we have some bound on the variance of a single random variable

$$\text{Var}_\mu[g] \leq \mathbb{E}[|\nabla g|^2]$$

and by tensorization, we can take the multivariate function f and derive

$$\text{Var}_\mu[f] \leq \mathbb{E}[||\nabla g||^2]$$

In here, we derive modified log-Sobolev inequalities by bounding the entropy of the form

$$\text{Ent}_\mu[e^g] \leq \mathbb{E}[|\nabla g|^2 e^g]$$

and then using tensorization to bound

$$\text{Ent}_\mu[e^{\lambda f}] \leq \mathbb{E}[||\nabla(\lambda f)||^2 e^{\lambda f}]$$

Lemma 8.6 (Discrete Modified log-Sobolev)

Let $D^- f \coloneqq f - \inf f$. Then,

$$\text{Ent}[e^f] \leq \text{Cov}[f, e^f] \leq \mathbb{E}[|D^- f|^2 e^f]$$

Proof.

Note that $\log \mathbb{E}[e^f] \geq \mathbb{E}[f]$ by Jensen's inequality. Therefore,

$$\text{Ent}[e^f] = \mathbb{E}[fe^f] - \mathbb{E}[e^f] \log \mathbb{E}[e^f] \leq \mathbb{E}[fe^f] - \mathbb{E}[f]\mathbb{E}[e^f] = \text{Cov}[f, e^f]$$

To prove the second part, we have

$$\text{Cov}[f, e^f] = \mathbb{E}[(f - \mathbb{E}[f]))(e^f - \mathbb{E}[e^f])] \leq \mathbb{E}[(f - \inf f)(e^f - e^{\inf f})]$$

and since e^x is convex, the first-order condition gives

$$e^{\inf f} \geq e^f + e^f(\inf f - f) \implies e^f - e^{\inf f} \leq e^f(f - \inf f)$$

and substituting above gives the result.

Now, by defining the one-sided differences

$$D_k^- f(x) = f(x_1, \dots, x_n) - \inf_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n)$$
$$D_k^+ f(x) = \sup_z f(x_1, \dots, x_{k-1}, z, x_{k+1}, \dots, x_n) - f(x_1, \dots, x_n)$$

we can use the discrete modified log-Sobolev inequality on each of them and then tensorize to get the following.

Theorem 8.5 (Bounded Difference Inequality)

For all $t \geq 0$,

$$\mathbb{P}[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n) \geq t] \leq \exp\left(- \frac{t^2}{4||\sum_{k=1}^n |D_k^- f|^2||_\infty} \right)$$
$$\mathbb{P}[f(X_1, \dots, X_n) - \mathbb{E}[f(X_1, \dots, X_n) \leq -t] \leq \exp\left(- \frac{t^2}{4||\sum_{k=1}^n |D_k^+ f|^2||_\infty} \right)$$

whenever $X_1, \dots, X_n$ are independent. In particular, $f(X_1, \dots, X_n)$ is subgaussian with variance proxy $2||\sum_{k=1}^n |D_k f|^2||_\infty$, where $D_k f = \sup_z f - \inf_z f$.

8.4 Modified log-Sobolev Inequalities

Theorem 8.6 (Modified log-Sobolov Inequality)

Let P_t be a Markov semigroup with stationary measure μ. The following are equivalent:
1. $\text{Ent}_\mu[f] \leq c\mathcal{E}(\log f, f)$ for all f (modified log-Sobolev inequality).
2. $\text{Ent}_\mu[P_t f] \leq e^{-t/c}\,\text{Ent}_\mu[f]$ for all f, t (entropic exponential ergodicity).

Moreover, if $\text{Ent}_\mu[P_t f] \to 0$ as $t \to +\infty$, then

$$\mathcal{E}(\log P_t f, P_t f) \leq e^{-t/c}\mathcal{E}(\log f, f) \text{ for all } f, t$$

implies 1 and 2 above.

9 Lipschitz Concentration and Transportation Inequalities

9.1 Concentration in Metric Spaces

Recall what a Lipschitz function is.

Definition 9.1 (Lipschitz Function)

Let (X, d) be a matrix space. A function $f : X \to \mathbb{R}$ is called L-**Lipschitz** if $|f(x) - f(y)| \leq L\, d(x, y)$ for all $x, y \in X$. The family of all 1-Lipschitz functions is denoted $\text{Lip}(X)$.

Remember that given iid $X_1, \dots, X_n \sim N(0, 1)$, Gaussian concentration states that the random variable is $||||\nabla f||^2||_\infty$-subgaussian. But we can write it in an equivalent way in terms of a Lipschitz property.

Lemma 9.1 ()

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a C^1 function. Then, $||||\nabla f||^2||_\infty \leq L^2$ if and only if f is L-lipschitz.

Therefore, if given random vector $X \sim N(0, I)$, then $f(X)$ is 1-subgaussian for every $f \in \text{Lip}(\mathbb{R}^n, ||\cdot||)$.