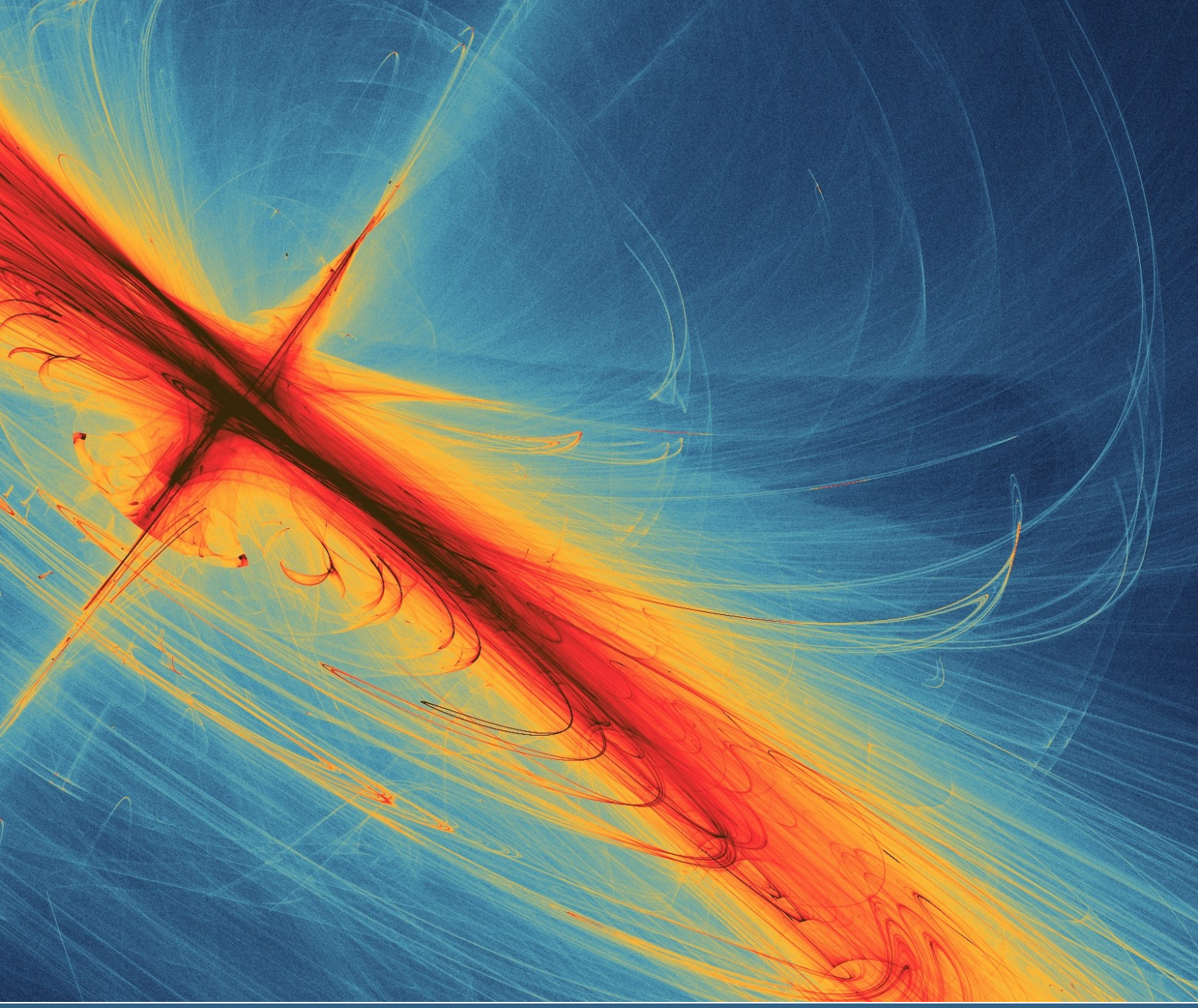


Notes: Spring 2021

Data Science, Stochastic Analysis, & PDEs

Author: Unique Divine

Institute: Columbia University

Date: Spring 2021

Contents

Part I

Stochastic Analysis

Chapter 1 Probability and Stochastic Processes

1.1 Motivation for learning about stochasic processes

The modelling of stochastic processes is one of the main applications of machine learning. A few examples:

- Poisson processes: For dealing with waiting times and queues.
- Random walk and Brownian motion processes: Used in algorithmic trading
- Markov decision processes: Commonly used in computational biology and reinforcement learning. HMMs are generally useful for understanding sequences and have applications for both writing and speech processing tasks.
- Auto-regressive and moving average processes: For time series analysis. ARIMA models.

1.2 Probability Theory I [Lec. 0, Jan 11]

1.2.1 Course Overview

The course will be non-traditional. It's not going to be your typical course found in a statistics or pure math department. What we'll do is present tools from stochastic analysis that are often useful in research and in industry for modeling physical systems.

The usual treatment of this subject is to go over some theoretical results and then talk about a few applications in finance. What we want to look at is the applications of this field in applied math. For instance, elliptic partial differential eqs, monte carlo methods, etc. This will cover the first few chapters of the textbook.

The goal is to gain an overall intuition for the subject, so we're not going to talk about all of the technical details. This doesn't mean we'll have fallacies in all of our derivations. It just means that we won't prove everything so that we can save time. We'll mostly look at the big picture and the connection to different things. We'll talk about why certain abstract things are actually useful.

Consequently, you'll see a lot of jumps. We'll also review elementary knowledge in this area and computing.

The first few homeworks will be a recap of some probability theory that we'll use. Limiting theorems, random variables, and distributions. The rest of the homework is mostly on projects. We will sometimes have simple derivations, schemes, code implementations of course concepts.

Today won't even be a review. We'll just mention what knowledge you will need.

1.2.2 Probability Theory Review

1. probability spaces: $(\Omega, F, \mathbb{P})$ = sample space, σ-algebra, probability measure

   A sigma alebra has a few properties (in first few chapters of textbook). "countable union"

   - $\phi \in F$
   - $A \in F \implies A^c \in F$
   - $\{A_i\}_{i=1}^{\infty} \in F \implies \cup A_i \in F$

   A probability measure is a function that maps between 0 and 1. $\mathbb{P} : f \to [0, 1]$.

   - $E_1 \subseteq E_2 \implies \mathbb{P}(E_1) \leq \mathbb{P}(E_2)$
   - Boole's Inequality: $\mathbb{P}(\bigcup_{i=1}^{\infty} E_1) \leq \sum_i \mathbb{P}(E_i)$
   - Inclusion-Exclusion: $\mathbb{P}(E_1 \cup E_2) = \mathbb{P}(E_1) + \mathbb{P}(E_2) - \mathbb{P}(E_1 \cap E_2)$

2. Conditional Probability
   - independence def.
   - conditional prob. def., Bayes' Thm
   - Law of total prob.: Let $\{E_i\}$ be pairwise disjoint s.t. $\bigcup_i E_i = \Omega$ and $\mathbb{P}(E_i) > 0$. Then, $\mathbb{P}(E) = \sum_i \mathbb{P}(E|E_i)\mathbb{P}(E_i) = \sum_i \mathbb{P}(E \cap E_i)$.

3. **Random Variables**.

   A random variable is a measurable real-valued function, $X(\omega) : \Omega \to \mathbb{R}$.

   Measurable $\equiv \ \forall x, \ \{\omega | X(\omega) \leq x\} \subset F$

   Distribution: The probability distribution function, $\mathbb{P}(X \leq x) = F_X(x)$. If

   $$\exists f_X(x) \text{ s.t. } F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \ \ \forall x,$$

   then f_X is a PDF and F_X is a CDF.

   Expectation: $\mathbb{E}[X] = \int_{\Omega} x f_X(t)dt$. Sometimes we write this more simply as

   $$\mathbb{E}[X] = \int_{\Omega} X d\mathbb{P} = \int_{-\infty}^{\infty} x f_X(x)dx.$$

   Thm: $X \geq 0 \implies \mathbb{E}[X] \geq 0$.

$$\mathbb{E}[a + bX] = a + b\mathbb{E}[X]$$

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

$$\{X_i\}_i \text{ independent} \implies \mathbb{E}[\prod_i X_i] = \prod_i \mathbb{E}[X_i]$$

Variance: Var(X) = $\mathbb{E}[(X - \mathbb{E}[X])^2]$

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$$

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}}$$

Also note that $\sigma_X^2 = \mathbb{E}[X^2] - \mathbb{E}[X]^2$

$$\text{Var}(a + bX) = b^2 \text{Var}(X)$$

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X,Y)$$

1.3 Probability Theory II [Lec. 1, Jan 13]

Moment inequalities

Thm Markov's Ineq

If $\mathbb{E}[X] < \infty$, then

$$\mathbb{P}(|X| \geq a) \leq \frac{\mathbb{E}[|X|]}{a}, \quad a \geq 0.$$

Proof

theorem: ϕ is monotone increasing

$$\mathbb{P}(|x| \geq a) = \frac{\mathbb{E}[\phi(|x|)]}{\phi(a)}.$$

Take $\phi(x) = x^2$.

$$\implies Y = |x - \mathbb{E}[x]|$$

$$\implies \mathbb{P}(|x - \mathbb{E}[x]| \geq a) \leq \frac{\mathbb{E}(|x - \mathbb{E}[x]|^2)}{a^2}$$

Why si this useful? It means that if you knwo how to control the variance, then you know how to control the probability. In the more general case, ϕ might be the third (or other higher order) moments.

Proof $\mathbb{P}(|x| \geq a) = \mathbb{P}(\phi(|x|) \geq \phi(a))$. Then Markov's Inequality.

Chebyshev Inequality is one of the fundamental inequalities you should have seen. You should also be familiar with moment generating functions.

Another one you should know: Jensen's Inequality.

Jensen's Inequality (Theorem): Let $f(x)$ be convex. Then, $\mathbb{E}[f(x)] \geq f(\mathbb{E}[x])$.

Another one that is important is cauchy-schwarz .

Cauchy Schwarz Inequality (Theorem): Suppose you have two random variables, X and Y s.t. $\mathbb{E}[X^2] < \infty$ and $\mathbb{E}[Y^2] < \infty$.

$$\implies \mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2].$$

Proof $\forall a, b \in \mathbb{R}$ define $Z = aX - bY$. You can then show that

$$\mathbb{E}[Z^2] = \mathbb{E}[(aX - bY)^2] = a^2\mathbb{E}[X^2] - 2ab\mathbb{E}[XY] + b^2\mathbb{E}[Y^2] \geq 0.$$
$$\implies \quad (2b\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2] \cdot b^2\mathbb{E}[Y^2] \leq 0$$
$$\implies \quad (\mathbb{E}[XY])^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

7. Characteristic Function

We're concerned with the characteristic fn of random variables, function spaces, or distributions. It's all the same stuff. It doesn't matter.

Let X be a R.V. on $(\Omega, F, \mathbb{P})$. Given $\phi(t) := \mathbb{E}[e^{itX}] \forall t \in \mathbb{R}$.

Note *This is called a fourier transform. It looks similar to the moment generating function,* $M_X(t) \equiv \mathbb{E}[e^{tX}], \quad t \in \mathbb{R}$.

$\phi(t) = \int_{-\infty}^{\infty} e^{itx} f(x)dx, \quad f(x)dx := dF(x)$

Example 1.

$X \sim \text{Unif}(a, b)$.

$\phi_X(t) = \mathbb{E}[e^{itX}] = \int_a^b$

Example 2.

$X \sim \mathcal{N}(0, 1)$.

$\phi_X(t) = \mathbb{E}[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$

$= e^{\frac{-1}{2}t^2}$

THm:

$$\phi(0) = 1 \tag{1.1}$$
$$|\phi(t)| \leq 1, \forall t \in \mathbb{R} \tag{1.2}$$

Proof

$$|\phi(t)| = |\int e^{itX} dF| \leq int|e^{itX}|dF \leq 1.$$

Thm: Let $\{x_k\}_{k=1}^n$ be independent. Let $z = \sum_k x_k$.

How do I find a distribution of z? We do a convolution.

$\phi_z(t) = \phi_{x_1} + \ldots + \phi_{x_n}$. By performing an inverse Fourier transform of the RHS, I can find the characteristic function. The "convolution" will give me the distribution. "We don't need to prove this."

Thm

Let x (from above) be s.t. $\mathbb{E}[x^n] < \infty$. Then, $\forall k \leq n$, $\phi^{(k)}(t) = i^k \int x^k e^{itx} dF(x)$

$\implies \phi^{(k)}(0) = i^k \int x^k dF(x) = i^k \mathbb{E}[x^k]$

$\implies \mathbb{E}[x^k] = i^{-k}\phi^{(k)}(0)$. The superscript notation denotes the kth derivative.

1.3.1 Law of Large Numbers (LLN)

Bernoulli's Weak LLN (Thm)

Why is it weak? We'll explore this. It has to do with weak convergence.

This theorem involves looking at a sequence of i.i.d. random variables. Let $\{x_n\}_{n\in N}$ be a seq of i.i.d. R.V.s with $\sigma^2 = \text{Var}(x_n)$

Define $S_n = \sum_{k=1}^{n} x_k$. Then, $\frac{S_n}{n} \xrightarrow{\mathbb{P}} \mu := \mathbb{E}[x_n]$ as $n \to \infty$.

definition of "convergence in probability"

$$\forall \epsilon > 0, \quad \lim_{n\to\infty} \mathbb{P}(|\frac{S_n}{n} - \mu| \geq \epsilon) = 0$$

Proof By the Chebyshev Ineq., $\mathbb{P}\left(|\frac{S_n}{n} - \mu| \geq \epsilon\right) \leq \frac{\mathbb{E}[(\frac{S_n}{n}-\mu)^2]}{\epsilon^2} = \frac{\frac{1}{n^2}\mathbb{E}[(S_n - n\mu)^2]}{\epsilon^2}$
$= \frac{textVar(S_n)}{n^2\epsilon^2} = \frac{n\sigma^2}{n^2\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \to 0.$

Kinchtin Weak LLN (Thm):

Let $\{X_n\}$ be i.i.d. be R.V. with $\mu := \mathbb{E}[X_n] < \infty$. Then, $\forall \epsilon, n \to \infty \implies \mathbb{P}(|\frac{S_n}{n} - \mu| \geq \epsilon) \to 0$.

Note *There is a final project, and you'll have more information about it throughout the next few weeks.*

A homework will come out next week on Monday. There's a link on courseworks to the office hours and the syllabus section.

1.4 Convergence of RVs I [Lec. 2, Jan 20]

1.4.1 Probability Theory III

Thm. Kolmogorow Strong Law of Large Numbers

Let $\{X_n\}$ be i.i.d. RVs with $\mathbb{E}[X_n] = \mu < \infty$. Then,

$$\mathbb{P}\left(\lim_{n\to\infty} \frac{S_n}{n} = \mu\right) = 1.$$

In real analysis, when do we say that a sequence of numbers converges?

$$\forall \epsilon > 0, \exists N \text{ s. t. } |x_n - x| \leq \epsilon \ \ \forall n \geq N.$$

Central Limit Theorem

If we take a random sample and have convergence, how fast will we see convergence? This is given by the Central Limit Theorem.

Let $\{X_n\}$ be i.i.d. with

$$\mathbb{E}[X_n] = \mu,$$
$$\text{Var}(X_n) = \sigma^2,$$
$$S_n = \left(\sum_{k=1}^{n} \frac{(X_k - \mu)}{n}\right) \frac{\sqrt{n}}{\sigma} = \frac{1}{\sqrt{n}} \sum_{k=1}^{n} \frac{(X_k - \mu)}{\sigma}.$$

Then, the sum will be a Guassian RV.

$$\lim_{n\to\infty} \mathbb{P}(S_n \leq x) = \Psi(X),$$

where $\Psi(X)$ is the CDF of $\mathcal{N}(0,1)$. The speed of convergence is $\frac{1}{\sqrt{n}}$.

Proof Calculate the characteristic function.

$$\phi_{S_n}(t) = \mathbb{E}[e^{it} S_n] = \mathbb{E}\left[\prod_{k=1}^{n} \exp\left(\frac{it}{\sqrt{n}}\left(\frac{X_n - \mu}{\sigma}\right)\right)\right]$$
$$= \prod_{k=1}^{n} \phi\left(\frac{t}{\sqrt{n}\sigma}\right) = \phi^n\left(\frac{t}{\sqrt{n}\sigma}\right)$$

$$\because \phi'(t) = \mathbb{E}[i(X_k - N)e^{it()}], \therefore \phi(0) = 1.$$
$$\phi'(0) = 0$$
$$\phi''(0) = -1$$

Taylor expand

$$= (1 - \frac{t^2}{2n\sigma^2} + O(n^2))$$
$$\phi_{S_n}(t) \xrightarrow{n\to\infty} e^{\frac{-t^2}{2}}$$

The above concludes the recap of what you are assumed to know from a previous probability course.

1.4.2 Convergence

Def [Convergence in Law/Distribution]: $\{X_n\}$ converges to X in law (or in distribution) if

$$\lim_{n\to\infty} F_n(X) = F(x), \ \forall\{x | x \text{ is continuous}\}.$$

Notation: $X_n \xrightarrow{D} X$

This is an extremely weak type of convergene.

Def [Convergence in Probability]: $\{X_n\}$ converge in probability to ... if

$$\forall \epsilon > 0, \ \lim_{n\to\infty} \mathbb{P}\left(|X_n - X| > \epsilon\right).$$

Notation: $X_n \xrightarrow{\mathbb{P}} X$.

Almost Sure Convergence (Def): This is also called convergence w/ prob 1. $\{X_n\}$ converges to X almost surely if

$$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1.$$

For each realization, you draw a sequence. Very strong. This is the convergence you see in terms of numbers. Notation: $X_n \xrightarrow{\text{a.s.}} X$. There's also something called "sure convergence", but we won't worry about it.

Convergence in ℓ^p norm (Def): AKA convergence in mean. $\{X_n\}$ converges to X in the ℓ^p norm if

$$\lim_{n\to\infty} \mathbb{E}\left[|X_n - X|^p\right] = 0.$$

For this definition to make sense, we require $\mathbb{E}[|X_n|^p] < \infty$. When $p = 1$, is it called convergence in mean. $p = 2 \implies$ convergence in mean-square.

(Thm):

- $X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X$
  and $\implies X_n \xrightarrow{D} X$.
- $X_n \xrightarrow{\ell^p} X \implies X_n \xrightarrow{\mathbb{P}} X$.
- $X_n \xrightarrow{\ell^p} X \implies X_n \xrightarrow{\ell^q} X, \quad 1 \le q \le p$.

Chapter 2 Markov Processes

When elements of a set are classified as being in one of several fixed states that can switch over time, this process is generally called a **stochastic process**. The swich between states in a stochastic process is described by a probability that, in general, depends on the current and previous states and the time in question [**?**].

- Ex.: An American voter's preference of political party could be the state. Voting cycles could be modeled as a stochastic process based on these preferences that change with time.

Stochastic prcoesses, i.e. random functions of time, are defined on a set, called the **index set**. The index set can be discrete or continuous. Markov chains have a discrete index set, while Poisson and diffusion process have continuous ones.

General stochastic processes are too broad of a class of objects to discuss in much detail, so we'll have to study specific cases of them instead. The specific case that will be the focus of this chapter is the Markov process.

If the probability to switch between two states of a stochastic process depends only on the two states in question (and not on the time, earlier states, or other factors), then this stochastic process is called a **Markov process**. To go a step further, if the number of possible states in the Markov process is finite, then the process is called a **Markov chain** [1][**?**]. Said another way, a Markov process is a process in which knowing the present state makes the future state(s) independent of the past.

http://langvillea.people.cofc.edu/MCapps7.pdf

https://math.libretexts.org/Bookshelves/Applied_Mathematics/Book%3A_Applied_Finite_Mathematics_(Sekhon_and_Bloom)/10%3A_Markov_Chains/10.02%3A_Applications_of_Markov_Chains

2.1 [Lec. 3, Jan 25]

Formal statment of a stochastic process: $X_t(\omega), t \in \boldsymbol{T}$. $\boldsymbol{T}$ is $\mathbb{R}$ or $\mathbb{N}$. Hence, $X_t(\omega) : \Omega \to \mathbb{R}$ or $X_t(\omega) : T \to \mathbb{R}$. In the latter case, it is called a trajectory or sample path. Again, this is a very wide set of functions, so we'll restrict our focus to something more specific to build intuition.

Markov Chain (Def): Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the probability space. A Markov chain is a sequence of random variables, $\{X_t\}_{t \in \boldsymbol{T}}$, parameterized with index set, $\boldsymbol{T}$, with the Markov property. Each X_t takes values in the state space, S.

[1]Notes will based on chapter 3 of Applied Stochastic Analysis by Weinan et al.

2.1.1 Discrete time finite Markov chains

Discrete in time means the parameterization is on a discrete set. Said another way, any countable set can be mapped onto it in a one-to-one manner.

Markov chain (Def): $\{X_n\}_{n\in\mathbb{N}}$ is a Markov chain if

$$\mathbb{P}(X_{n+1} = x_{n+1}|\{X_k = x_k\}_{k=1}^n) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n).$$

Intuitive definition: The next state is only dependent upon the current state.

Ex. 1 - Markov chain:

$$\zeta_k := \text{ i.i.d. R.V. s.t. } \begin{cases} 1 & \mathbb{P} = 0.4 \\ -1 & \mathbb{P} = 0.6 \end{cases}$$

$$X_n := \sum_{k=1}^{n} \zeta_k \text{ is then Markovian.}$$

$$\implies X_{n+1} = \sum_{k=1}^{n+1} \zeta_k = X_n + \zeta_{n+1}.$$

$$\mathbb{P}(X_{n+1} = x_{n+1}|\{X_k = x_k\}_{k=1}^n) = \mathbb{P}(X_{n+1} = x_{n+1}|X_n = x_n)\mathbb{P}(X_{n+1} = X_n + 1|X_n = x_n) = 0.4$$

$$\mathbb{P}(X_{n+1} = X_n - 1|X_n = x_n) = 0.6$$

Having both probabilities set to 0.5 is called symmetric random walk.

Finite Markov chain (Def): A Markove chain is finite if its state space is finite. In Ex. 1, the Markov chain is not finite because its state space is infinite and countable. You could make it finite by taking the modulus and condensing the state space.

Ex. 2 - Boolean Stock Market:

$$X_n = \begin{cases} 1 & \text{bull year} \\ -1 & \text{bear year} \end{cases}$$

$$\mathbb{P}(X_n = 1|X_n = 1) :=$$

$$\mathbb{P}(X_n = -1|X_n = -1) := 0.4$$

TODO: (above)

Chapman-Kalmogorov Eq. (Thm):

Let $\{X_n\}$ be a Markov chain starting in state, $X_0 = i$. Assume the state space, S, is countable. Then,

$$\mathbb{P}(X_n = j|X_0 = i) = \sum_{k\in S} \mathbb{P}(X_n = j|X_m = k)\mathbb{P}(X_m = k|X_0 = i), \quad \forall 1 \leq m \leq n-1.$$

But what does this mean? Suppose you have a countable sequence of events, $\{E_k\}_{k=1}^{\infty}$ s.t. $\bigcup_{k=1}^{\infty} E_k = \Omega$ and $E_k \cap E_{k'} = \phi, \forall k, k'$. And, $\mathbb{P}(F) = \sum_{k=1}^{\infty} \mathbb{P}(F \cap E_k)$. So, Chapman-Kalmogorov is basically the law of total probability twisted a bit.

You have a process in which you're jumping from state i to state j. This theorem states that the probability of such an setup is the sum of all possible intermediate jumps.

Proof

$$\begin{aligned}
\mathbb{P}(X_n = j|X_0 = i) &= \sum_{k\in S} \mathbb{P}\left(X_n = j \cap X_m = k|X_0 = i\right) \\
&= \sum_{k\in S} \mathbb{P}\left(X_n = k|X_m = k \cap X_0 = i\right) \mathbb{P}(X_m = k|X_0 = i) \\
&= \sum_{k\in S} \mathbb{P}(X_n = j|X_m = k)\mathbb{P}(X_m = k|X_0 = i) \quad \text{(Markov assumption)}
\end{aligned}$$

Invariant distribution of stationary Markov chains: Let $S = \{1, 2, \ldots, I\}$ be a countable set of states. Don't be alarmed by these integers in S. These are just labels for the states similar to how we labeled bear and bull markets 1 and -1 in Ex. 2. With this state, we can define the transition probability.

Transition probability (Def): The transition probability at step n is

$$\mathbb{P}_{kj}^{(n)} = \mathbb{P}(X_{n+1} = j|X_n = k).$$

The superscript means "at step [superscript]". The order of the symbols in the subscript indicates the order of events, so you may see $\mathbb{P}_{kj}^{(n)} = \mathbb{P}(X_n = j|X_0 = k)$ to mean the same thing in another text. In our notation, it means "transition from state k to j." If $\mathbb{P}_{kj}^{(n)}$ is independent of n, we say that the MC is **stationary**.

Stationary transition matrix: With a stationary transition probability $\mathbb{P}_{kj}$, a stationary transition matrix can be defined as $P = (P_{kj})_{kj\in S}$. Its columns are probability vectors.

- $P_{kj} \geq 0 \; \forall k, j.$
- $\sum_{j\in S} P_{kj} = 1 \;\; \forall k \in S.$

2.2 [Lec. 6, Feb. 3]

We have some observations, $\{Y_n\}$, in the HMM. See Kui Ren's Lecture Notes 01-C.pdf.

An Emmission matrix of a HMM is R with elements $R_ij = Y_ij$, where $i \in S$ and $j \in O$. $Y_ij := \mathbb{P}(Y_j = y|X_i = x)$ The parameters for this model are $\theta := (\boldsymbol{\mu}_0, \boldsymbol{P}, \boldsymbol{R})$. The main quantities of interest for the two sequences of random variables $X = (X_1, \ldots, X_n)$, $Y = (Y_1, \ldots, Y_n)$:

- $\mathbb{P}(X|\theta) = \boldsymbol{\mu}_{0,X_1}\mathbb{P}_{X_1X_2}\mathbb{P}_{X_2X_3}\cdots\mathbb{P}_{X_{n-1}X_n} = \boldsymbol{\mu}_{0,X_1}\prod_{k=1}^{n-1}\mathbb{P}_{X_kX_{k+1}}$
- By the law of total probability, $\mathbb{P}(Y|X,\theta) = \mathbb{P}(Y_1|X_1,\theta)\mathbb{P}(Y_2|X_2,\theta)\cdots\mathbb{P}(Y_n|X_n,\theta) = \prod_{k=1}^{n}\mathbb{P}(Y_k|X_k,\theta)$
- $\mathbb{P}(X,Y|\theta) = \mathbb{P}...$

2.2.1 Parameter Estimation in HMMs

: We'd like to know the mapping of $Y \to \boldsymbol{\theta}$ in order to estimate the parameter $bm\theta$.

Q: How do you find the parameter $\boldsymbol{\theta}$ in a HMM?

We know $\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta})$, which is the parameter-to-observation map, and can use maximum likelihood methods to find the parameter, $\boldsymbol{\theta}$, that most likely has the generated observation, $\boldsymbol{Y}$. In other words, we solve for $\mathbb{P}(\boldsymbol{\theta}|\boldsymbol{Y}) = \dfrac{\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta})\mathbb{P}(\boldsymbol{\theta})}{\mathbb{P}(\boldsymbol{Y})}$.

Q: How does a maximum likelihood method work?

Method 1: The parameter that has the largest chance to generate the observation is $\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{theta}}\mathbb{P}(\boldsymbol{Y}|\boldsymbol{\theta})$. It is then called the maxmimum likelihood estimator.

Method 2: Maximum a posteriori (MAP) estimation: A Bayesian approach.

2.2.2 Continuous Time Finite Markov Chain

Lecture Notes 01-D.pdf

Let $\{X_t\}_{t\in\mathbb{R}^+}$ be a continuous time stochastic process. In order for $\{X_t\}$ to be a Markov process, it must satisfy the Markov property:

$$\mathbb{P}(X_{t+s} = x_{t+s}|\{X_{t'}\}) = \mathbb{P}(X_{t+s} = x_{t+s}|X_s = x_s),$$

where $t' \in [0, s]$.

- $\{X_t\}$ is right-continuous. $\therefore \lim_{h\to 0_+} X_{t+h} = X_t$.
- $\{X_t\}$ has a finite state space, $S = \{1, 2, \cdots, I\}$.
- Transition probability: $p_{jk}(t) = \mathbb{P}(X_{t+s} = k|X_s = j)$
- Assumed stationarity, or homogeneity in time. This means that the transition probability is independent of s.

Q: This is the continous picture for finite Markov Chains. How do you map to the discrete case?

The discrete Markov chain is simply the continuous case fixed with $t = 1$ in the transition probability. Said another way, the step size of the "jump" has size 1 for the discrete case rather than size t.

Notes from homework 2:

- problem 1 on bottom of page 6
- Convergence proofs
- Convergence in prob. of the sum of two RVs
- More convergence proofs
- Even more convergence proofs- Q4 is on pages 3-4.

Chapter 3 Feb.

3.1 Homework 3

3.1.1 Q 7 - Brownian Motion

Definition 3.1. Brownian motion / Wiener process

A Wiener process is a stochastic process $\{W_t\}_{t\geq 0}$ with three properties: continuity of path, normality of increment, and independent increment.

(cloze) In Brownian motion, normality of increment means that $\forall t > s \geq 0, W_t - W_s \sim \mathcal{N}(0, t-s)$.

(cloze) In Brownian motion, independence of increment means that $\forall t \geq s \geq 0, W_t - W_s$ is independent of $W_{s'}$ $\forall\{s' \mid 0 \leq s' \leq s\}$.

(cloze) TODO (continuity of path)

♣

Q: (cloze) A Wiener process is also known as a Brownian motion.

(cloze) A Brownian motion is type of stochastic process.

(cloze) A Brownian motion is said to be standard if $W_0 = 0$.

Q: Why is a Wiener process, $\{W_t\}_{t\geq 0}$, called standard if $W_0 = 0$.

This follows from the normality of increment property: $\forall t > s \geq 0, W_t - W_s \sim \mathcal{N}(0, t-s)$. Thus,

$$W_0 = 0 \implies W_t - W_s\Big|_{s=0} = W_t - W_0 = W_t \sim \mathcal{N}(0,t).$$

In other words, W_t has a standard normal distribution.

Fact 3.1. Standard Brownian motion covariance

$\text{Cov}(B_t, B_s) = \min(s,t)$

♠

3.1.2 Q 6 - Infitestimal generator

3.1.3 Q 2 - Metroplis Hastings Explained

Awesome notes

Notes from homework 3:

- problem 1 on bottom of page 6
- Columba IEOR 4700 Brownian notes
- Advanced Mathematical Finance Hwk solutions
- Karl Sigman - Notes on Stochastic Modeling I
- Karl Sigman - Notes on Simulation
- Matrix derivatives

- Recommended books on stochastic processes
- Markov Chains - Illinois

Chapter 4 March - Stochastic Analysis

4.1 Homework 4

Definition 4.1. Covariance of r.v.s

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}[Y])]$$
$$\text{Var}(X) = \text{Cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

♣

(cloze) Two r.v.'s X and Y are called uncorrelated if $\text{Cov}(X,Y) = 0$.

Q1 | Standard Wiener process

Let $\{W_t\}_{t\geq 0}$ be a standard Wiener processes. Evaluate the following quantities.

(i) $\mathbb{E}[W_t^4]$.

(ii) $\mathbb{E}[(W_t - W_s + W_z)^2], \quad t,s,z \in [0,1]$.

Q1 i

Here, I'm tasked with finding the expectation of a Brownian motion. Expectiations are defined by $\mathbb{E}(f(X)) = \int f(s)\rho_X(s)\mathrm{d}s$, where ρ_X is the PDF of r.v. X.

So, if we wanted to compute $\mathbb{E}[W_t^4]$, this would be

$$\mathbb{E}[f(W_t)] = \int f(w)\rho_{W_t}(w)\mathrm{d}w$$
$$\mathbb{E}[W_t^4] = \int w^4\rho_{W_t}(w)\mathrm{d}w \tag{1}$$

What then is ρ_{W_t}? Well, we know that $\{W_t\}$ is a standard Brownian motion. Thus by the normality of increment property for a Brownian motion, $W_t - W_s \sim \mathcal{N}(0, t-s)$ for all $t > s \geq 0$. Since the motion is standard, $W_0 = 0$, which implies that $W_t - W_0 = W_t \sim \mathcal{N}(0,t)$ for all $t > 0$.

$$W_t \sim \mathcal{N}(0,t) \iff \rho_{W_t}(w) = \frac{1}{\sqrt{2\pi t}}\exp\left(\frac{-w^2}{2t}\right) \tag{2}$$

Part (i) can be completed by computing the integral in Eq. 1 using the PDF (Eq. 2).

Q1 ii

Computing $\mathbb{E}[(W_t - W_s + W_z)^2], \quad t,s,z \in [0,1]$ isn't as straightforward.

Q: Are the W_s and W_z terms still standard, and how could I tell?

Yes, defining $\{W_t\}_{t\geq 0}$ as a standard Weiner process with $t \in [0,1]$ implies that W_s and W_z would just denote $W_t\big|_{t=s}$ and $W_t\big|_{t=z}$, respectively (as long as $s,z \in [0,1]$).

Q: Derive the expectation of a standard Wiener process.

These follow directly from the definitions. If $\{B_t\}$ is a standard B.M. , then $B_0 = 0$ and $B_t - B_s \sim \mathcal{N}(0, t-s)$ for $0 \leq s < t \in [0,1]$. Thus, it's clear from $\mathbb{E}[B_t - B_s] = 0$ that $\mathbb{E}B_t = \mathbb{E}[B_t - B_0] = 0$ too.

Q: What is the expecation of a nonstandard Wiener process?

TODO

(i) $\{X_t\}_{t\geq 0}$ with $X_t := W_t - tW_1$ is a Brownian bridge.

$$
\begin{aligned}
m(t) = \mathbb{E}X_t &= \mathbb{E}[W_t - tW_1] \\
&= \mathbb{E}W_t - t\mathbb{E}W_1, \quad W_t \sim \mathcal{N}(0,t), W_1 \sim \mathcal{N}(0,1) \\
\therefore \quad &\boxed{m(t) = 0} \\
K(s,t) = \text{Cov}(X_t, X_s) &= \text{Cov}(W_t - tW_1, W_s - sW_1) \\
&= \text{Cov}(W_t, W_s - sW_1) - t\text{Cov}(W_1, W_s - sW_1) \\
&= \text{Cov}(W_t, W_s) - s\text{Cov}(W_t, W_1) - t\text{Cov}(W_1, W_s) \\
&\quad + st\text{Cov}(W_1, W_1) \\
&= \min(s,t) - s\min(t,1) - t\min(1,s) + st\min(1,1) \\
&= \min(s,t) - st - st + st \\
\therefore \quad &\boxed{K(s,t) = \min(s,t) - st}.
\end{aligned}
$$

$\{X_t\}$ fits the definition for a Brownian bridge.

(ii) $\{Y_t\}_{t\geq 0}$ with $Y_t := (1-t)W_{t/(1-t)}$ for $0 \leq t < 1$, $Y_1 = 0$ is a Brownian bridge.

Definition 4.2. Brownian bridge

A Brownian bridge $\{X_t\}$ is a Gaussian stochastic process s.t. $X_t \equiv B_t - tB_1$, where $\{B_t\}_{t\geq 0}$ is a standard Brownian motion. ♣

Fact 4.1. Brownian bridge covariance

Let $\{X_t\}_{t\geq 0}$ be a Brownian bridge defined by $X_t := W_t - tW_1$.

$$\begin{aligned} K(s,t) = \text{Cov}(X_t, X_s) &= \text{Cov}(W_t - tW_1, W_s - sW_1) \\ &= \text{Cov}(W_t, W_s - sW_1) - t\text{Cov}(W_1, W_s - sW_1) \\ &= \text{Cov}(W_t, W_s) - s\text{Cov}(W_t, W_1) - t\text{Cov}(W_1, W_s) + st\text{Cov}(W_1, W_1) \\ &= \min(s,t) - s\min(t,1) - t\min(1,s) + st\min(1,1) \\ &= \min(s,t) - st - st + st \\ \therefore \quad &\boxed{K(s,t) = \min(s,t) - st}. \end{aligned}$$

♠

Fact 4.2. Brownian bridge expectation

Let $\{X_t\}_{t\geq 0}$ be a Brownian bridge defined by $X_t := W_t - tW_1$.

$$\begin{aligned} m(t) = \mathbb{E}X_t &= \mathbb{E}[W_t - tW_1] \\ &= \mathbb{E}W_t - t\mathbb{E}W_1, \quad W_t \sim \mathcal{N}(0,t), W_1 \sim \mathcal{N}(0,1) \\ \therefore \quad &\boxed{m(t) = 0} \end{aligned}$$

♠

4.1.1 Q2 - Brownian Bridge

Prove a stochastic process $\{X_t\}$ is a Brownian bridge. This involves proving the process has mean function $m(t) = 0$ and covariance function $K(s,t) = \min(s,t) - st$ for $s,t \in [0,1]$.

Q: What is a Guassian process? "standard Brownian bridge is a Gaussian process with continuous paths..."

§5.4 Guassian Processes (E et al., 2020)

(cloze) A stochastic process $\{X_t\}_{t\geq 0}$ is called a Guassian process if its finite-dimensional distributions $\mu_{\{t_i\}_{i=1}^k}$ are consistent Gaussian measures for any $0 \leq t_1 < t_2 < \cdots < t_k$.

(cloze) A gaussian process $\{X_t\}$ is determined once its mean and covariance function,

$$m(t) = \mathbb{E}X_t \quad \text{and} \quad K(s,t) = \mathbb{E}\left[(X_s - m(s))(X_t - m(t))\right],$$

are specified.

(cloze) It is well known that a Guassian random vector $\boldsymbol{X} \in M_{n\times 1}$, where $X_i = \boldsymbol{X}_i$ are random variable components of $\boldsymbol{X}$, is completely characterized by its first and second moments,

$$\boldsymbol{m} = \mathbb{E}\boldsymbol{X} \quad \text{and} \quad \boldsymbol{K} = \mathbb{E}\left[(\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})^T\right].$$

In component form,

$$m_i = \mathbb{E}X_i \quad \text{and} \quad \boldsymbol{K}_{ij} = \mathbb{E}\left[(X_i - m_i)(X_j - m_j)\right].$$

(cloze) Using $\boldsymbol{m}$ and $\boldsymbol{K}$, one can represent $\boldsymbol{X}$ via its characteristic function,

$$\mathbb{E}e^{i\boldsymbol{\epsilon}\cdot\boldsymbol{X}} = \mathbb{E}e^{i\boldsymbol{\epsilon}^T\boldsymbol{X}} = e^{i\boldsymbol{\epsilon}^T\boldsymbol{m}-\frac{1}{2}\boldsymbol{\epsilon}^T\boldsymbol{K}\boldsymbol{\epsilon}}.$$

Q: What is a characteristic function in general?

Source: wikipedia - chacteristic function

(cloze) In probability theory and statistics, the charactristic function of any real-valued stochatic variable completely defines its probability distribution.

(cloze) If a random variable admits a probability density fn. (PDF), then the characteristic fn. is the Fourier transform of the PDF.

(cloze) The characteristic function always exists when treated as a function of a real-valued argument, unlike the moment-generating function.

(cloze) Similar to the cumulative distribution function (CDF), the characteristic fn. provides an alternative way to describe a stochastic variable.

(cloze) For random variable, X, the characteristic function is defined by $\phi_X(t) \equiv \mathbb{E}e^{itX}$.

(cloze) If a random variable admits a probability density fn. (PDF), then the characteristic function is its dual, which means that each of them is a Fourier transform of the other.

(cloze) If a rand.var. has a moment-generating fn., $M_X(t)$, then the domain of the characteristic fn. can be extended to the complex plane: $\phi_X(-it) = M_X(t)$. Note however that the characteristic fn. of a distribution always exists, even when the prob. density fn. and moment-generating fn. do not.

Q: How do we find covariance functions?

First, let's recall some definitions about variance and covariance. **(cloze)**

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}[Y])]$$
$$\text{Var}(X) = \text{Cov}(X,X) = \mathbb{E}[(X - \mathbb{E}[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2$$

(cloze) Two r.v.'s X and Y are called uncorrelated if $\text{Cov}(X,Y) = 0$.

All of these definitions extend to the vector case.

(cloze) If $\boldsymbol{X} \in \mathbb{R}^d$ is a stochastic vector s.t. each component $X_k \in \boldsymbol{X}$ is a random variable, then the covariance matrix of $\boldsymbol{X}$ is defined as

$$\text{Cov}(\boldsymbol{X}) = \mathbb{E}\left[(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})(\boldsymbol{X} - \mathbb{E}\boldsymbol{X})^T\right].$$

This is expectation of the dyadic (outer) matrix product of $\boldsymbol{X} - \mathbb{E}\boldsymbol{X}$ and itself.

4.1.2 Q3 - Karhunen-Loeve

Q3: Derive the Karhunen-Loeve expansion for the standard Brownian bridge. *Hint: you will need to find eigenpairs of the operator* $\mathcal{K}f := \int_0^1 K(s,t)f(s)ds$ *on* $L^2([0,1])$ *as what we did for the standard Brownian motion.*

Q: Karhunen-Loeve expansion?

page 112 of E et al., Thm. 5.13

4.1.3 Q4 - Generators

Q4: Find the generator of the standard Brownian bridge.

generator = infitestimal generator

4.1.4 Q5 - Fractional Brownian motion

Q5: A stochastic process $\{B_t^H\}_{t\geq 0}$ is called a **fractional Brownian motion** if it is a Gaussian process with mean $m(t) = 0$ and covariance

$$K(s,t) = \frac{1}{2}(t^{2H} + s^{2H} - |t-s|^{2H}), \quad s,t \in [0,T].$$

The parameter $H \in (0,1)$ is called the Hurst index. Prove that $\{B_t^H\}$ has the following properties:

...

1. (Self-similarity): $B_{\beta t}^H$ has the same distribution as $\beta^H B_t^H$ for any $\beta > 0$;
2. (Stationary increment): $B_t^H - B_s^H$ has the same distribution as B_{t-s}^H for $0 \leq s < t$.
3. $\{B_t^H\}_{t\geq 0}$ with $B_0^H = 0$ and $H = \frac{1}{2}$ is a standard Brownian motion.

"has the same distribution" $\rightarrow$ Is a good way to go about

Probability-generating function

Generating function

4.1.5 Reading Assignments

- Review §6.1-6.7 of E, Li & Vanden-Ejinden
- Review Lecture Notes 02 (D-E)

4.1.6 References - HW4

- problem 1 on bottom of page 6

Part II

PDE

Chapter 5 ODE Solving

5.1 Variation of Parameters

I'll assume knowledge of how to solve homogeneous ODE such as the following ones.

Q: $v = v(t)$**. Solve** $v'' + v = 0$**.**

$$\boxed{v(t) = c_0 \cos(t) + c_1 \sin(t)}.$$

Q: $v = v(t)$**. Solve** $v'' - v = 0$**.**

$$\boxed{v(t) = c_0 \cosh(t) + c_1 \sinh(t)}.$$

Q: How do we solve the more general $av''(t) + bv'(t) + c = f(t)$ **for a general** $f(t)$**?**

Method of variation of parameters

Without any proof for why the method works, I'll go over how to use it. First, I'll need to talk about Cramer's rule and Wronskians.

5.1.1 Cramer's Rule

Cramer's rule can be used to solve linear systems of equations that have a unique solution. We'll focus on systems of the following form since it will be relevant for variation of parameters:

$$Xc' = \beta$$
$$\begin{bmatrix} x_0 & x_1 \\ x_0' & x_1' \end{bmatrix} \begin{bmatrix} c_0' \\ c_1' \end{bmatrix} = \begin{bmatrix} 0 \\ f \end{bmatrix},$$

where everything is a function of the same variable (let's call it t).

Note *X is a called a "Wronskian" of x_0 and x_1.*

$$X = W(x_0, x_1) = \begin{bmatrix} x_0(t) & x_1(t) \\ x_0'(t) & x_1'(t) \end{bmatrix}$$

This system has a unique solution if and only if $\det(X) \neq 0$. Let's suppose that's true. Since $\det(X) = x_0x_1' - x_1x_0' \neq 0$, we can easily solve for $c = X^{-1}\beta$.

$$\begin{bmatrix} c_0' \\ c_1' \end{bmatrix} = \frac{1}{x_0x_1' - x_1x_0'} \begin{bmatrix} x_1' & -x_1 \\ -x_0' & x_0 \end{bmatrix} \begin{bmatrix} 0 \\ f \end{bmatrix} = \frac{1}{\det(W(x_0, x_1))} \begin{bmatrix} -x_1f \\ x_0f \end{bmatrix}$$

Equivalently, we could apply Cramer's Rule, which gives us the same answer from computing determinants. It's also requires less memorization.

$$c_0' = \frac{\begin{vmatrix} 0 & x_1 \\ f & x_1' \end{vmatrix}}{\det W(x_0, x_1)}, \quad c_1' = \frac{\begin{vmatrix} x_0 & 0 \\ x_0' & f \end{vmatrix}}{\det W(x_0, x_1)}.$$

Either way, we can solve for the desired coefficients since we know the value of their derivatives.

$$c_0(t) = \int_0^t c_0'(t)\mathrm{d}t + d_0$$

$$c_1(t) = \int_0^t c_1'(t)\mathrm{d}t + d_1$$

5.1.2 Method of undetermined coefficients

The method of undetermined coefficients involves making educated guesses about the form of the particular solution to an ODE based on the form of non-homogeneous portion.

A key pitfall of this method is that the form needed for the initial guess is not obvious. It's often better to use variation of parameters with Cramer's rule.

Q: Find the particular solution to $y'' + 4y' + 3y = 3x$.

The non-homogeneous component is a polynomial, so we assume particular solutions of polynomial form up to the order of the compoment, i.e. $Ax + B$.

$$y_p = Ax + B$$

$$\left(y'' + 4y' + 3y\right)\Big|_{y_p} = 0 + 4(A) + 3(Ax + B) = 3x$$

$$(3A)x + (4A + 3B) = 3x + 0$$

$$\therefore \quad 4A + 3B = 0 \text{ and } 3A = 3.$$

$$\therefore \quad A = 1, \quad B = -1.$$

$$\therefore \quad \boxed{y_p = x - 1}$$

Q: Find the homogeneous solutions to $y'' + 4y' + 3y = 3x$.

Assume exponential solutions of the form $y_h(x) = e^{mx}$:

$$(m^2 + 4m + 3)e^{mx} = 0$$

$$m^2 + 4m + 3 = 0 \qquad (e^{mx} \neq 0)$$

$$(m + 3)(m + 1) = 0$$

$$m = \{-3, -1\}.$$

$$\therefore \quad \boxed{y_h = c_0 e^{-3x} + c_1 e^{-x}}$$

Chapter 6 Heat Equation

6.1 Equilibrium Temperature distribution (Haberman §1.4)

The simple problem of heat flow:

Q: (cloze) If thermal coefficients are constant and there are no sources of thermal energy, then the temperature $u(x,t)$ in 1D rod $0 \le x \le L$ satisfies

$$\partial_t u = k\partial_x^2 u.$$

Q: (cloze) The above is known as the heat equation in 1D.

Q: What is the precise meaning of steady-state in relation to the heat equation?

If we say the boundary conditions at $x = 0$ and $x = L$ are steady, that means they are independent of time. $\implies$ We define an equlibirum or steady-state solution to the heat equation is one that does not depend on time, i.e. $u(\vec{x},t) = u(\vec{x})$.

Q: Solve $\partial_x^2 u = 0$.

$$\partial_x^2 u = 0 \implies \frac{\partial}{\partial x}(\partial_x u) = 0$$
$$\mathrm{d}(\partial_x u) = 0 \cdot \mathrm{d}x \implies \int d(\partial_x u) = \int 0 \cdot \mathrm{d}x$$
$$\therefore \partial_x u = C_0$$
$$\int \partial_x u \mathrm{d}x = \int C_0 \mathrm{d}x. \quad \therefore \quad \boxed{\partial_x^2 u = C_0 x + C_1}$$

Q: For equlibirium diffusion in a 1D rod with $x \in [0, L]$, what are the boundary conditions and constraints?

Equilibrium $\implies u = u(\vec{x}), \iff u(0,t) = T_0$ and $u(L,t) = T_1$.

Also, $\nabla^2 u = 0$.

Q: Determine the equilibrium temperature distribution for a 1-D rod ($x \in [0, L]$) with constant thermal properties with the following source and boundary conditions:

$$Q = 0,\ u(0) = 0,\ u(L) = T.$$

$$\text{PDE: } \partial_t u = k\nabla^2 u + Q \qquad \text{(heat eq)}$$

$$\text{ODEs (equilibrium): } k\nabla^2 u = 0.\ \ \partial_t u = 0$$

$$\nabla^2 u = 0 \implies u = c_0 x + c_1$$

$$u(0) = 0 \implies c_1 = 0$$

$$u(L) = T \implies c_0 = \frac{T}{L}.$$

$$\therefore \boxed{u(x,t) = \frac{T}{L}x}.$$

Q:

Lec. 3

How do we find B_n?

Fourier's trick. Multipl by

$$f(x) = \sum_{n=1}^{\infty} B_n \sin\left(\frac{n\pi}{L}x\right)$$

Orthogonality condition for sine

$$\int_0^L \sin\left(\frac{n\pi}{L}x\right)\sin\left(\frac{m\pi}{L}x\right)\mathrm{d}x = \begin{cases} \frac{L}{2} & m = n \\ 0 & m \neq n \end{cases}$$

$B_m = \frac{2}{L}\int_0^L f(x)\sin\left(\frac{m\pi}{L}x\right)\mathrm{d}x$

Example - Diffusion Eq.

Given:

$$\partial_t = k\partial_x^2 u \ \ t > 0, x \in (0, L)$$

$$u(0,t) = u(L,t) = 0$$

$$u(x,0) = 1$$

We just derived the solution to this general eq., which is

$$u(x,t) = \sum_{n=1}^{\infty} B_n e^{-k(\frac{n\pi}{L})^2 t}\sin\left(\frac{n\pi}{L}x\right).$$

$$B_n = \frac{2}{L}\int_0^L f(x)\sin\Big(\frac{n\pi}{L}x\Big)\mathrm{d}x = \frac{2}{L}\int_0^L \sin\Big(\frac{n\pi}{L}x\Big)\mathrm{d}x$$
$$= \frac{2}{L}\left(-\cos\Big(\frac{n\pi}{L}x\Big)\right)\Bigg|$$

Lec. 4

Recap:

We found that the 1D heat equation on a rod ($x \in (0, L)$), $\partial_t = k\partial_x^2 u$, subject to the following boundary conditions:

$$u(0,t) = u(L,t) = 0$$
$$u(x,0) = f(x)$$

has the solution

$$u = \sum_{n=1}^{m} B_n e^{-k(\frac{n\pi}{L})^2 t}\sin\Big(\frac{n\pi}{L}x\Big)$$
$$u(x,0) = f(x) = \sum_{n=1}^{m} B_n \sin\Big(\frac{n\pi}{L}x\Big)$$
$$B_n = \frac{2}{L}\int_0^L f(x)\sin\Big(\frac{m\pi}{L}x\Big)\mathrm{d}x = \begin{cases} 0 & \text{if } n \text{ is even} \\ \frac{4}{n\pi} & \text{if } n \text{ is odd} \end{cases}$$
$$\therefore u(x,t) = \sum_{n_{\text{odd}}\geq 1}^{\infty} \frac{4}{n\pi} e^{-k(\frac{n\pi}{L})^2 t}\sin\Big(\frac{n\pi}{L}x\Big)$$

§2.4: Heat conduction in a rod with insulated ends BCs: $\partial_x u(0,t) = \partial_x u(L,t) = 0$

IC:

Solutions of the form: $u(x,t) = A_0 + \sum_{n=1}^{\infty} A_n e^{-k(\frac{n\pi}{L})^2 t}\cos\big(\frac{n\pi}{L}x\big)$.

Let's find the arbitrary coefficients A_0 and A_n ($n \geq 1$).

$$u(x,0) = f(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos\Big(\frac{n\pi}{L}x\Big) \tag{6.1}$$

Q: Solve for A_0.

Integrate and interchange the order of sum and integral.

$$\int_0^L f(x)\mathrm{d}x = \int_0^L A_0\mathrm{d}x + \sum_{n=1}^{\infty} A_n \int_0^L \cos\Big(\frac{n\pi}{L}x\Big)\mathrm{d}x$$
$$= A_0 L + \sum_{n=1}^{\infty} A_n \left(\frac{L}{n\pi}\sin\Big(\frac{n\pi}{L}x\Big)\Big|_0^L\right)$$
$$= A_0 L$$

$$A_0 = \frac{1}{L}\int_0^L f(x)\mathrm{d}x \tag{6.2}$$

Q: Solve for A_n.

Fourier's Trick, i.e. multiply both sides by $\phi_m(x) = \alpha_1 \cos\left(\frac{m\pi}{L}x\right)$ and integrate.

$$\begin{aligned}\int_0^L f(x)\phi_m(x)\mathrm{d}x &= \sum_0^\infty A_n \int_0^L \phi_n(x)\phi_m(x)\mathrm{d}x \\ &= \sum_0^\infty A_n \int_0^L \cos\left(\frac{n\pi}{L}x\right)\cos\left(\frac{m\pi}{L}x\right)\mathrm{d}x\end{aligned}$$

We know from the orthogonality relations of cosine that this integral is 0 everywhere except for $m = n$. Consequently,

$$\int_0^L f(x)\phi_m(x)\mathrm{d}x = A_m \int_0^L \phi_m^2(x)\mathrm{d}x$$

$$A_m = \frac{\int_0^L f(x)\phi_m(x)\mathrm{d}x}{\int_0^L \phi_m^2(x)\mathrm{d}x} = \frac{\int_0^L f(x)\phi_m(x)\mathrm{d}x}{(\frac{L}{2})} = \frac{2}{L}\int_0^L f(x)\phi_m(x)\mathrm{d}x$$

Q:

$$\therefore \quad A_0 = \frac{1}{L}\int_0^L f(x)\mathrm{d}x. \tag{6.3}$$

Diffusion eq. in an insulated circular ring [Lec. 4, 1-21]

BCs: Periodic boundary conditions

Chapter 7 Laplace's Eq.

- Book §2.5
- Start: Lecture 4, 1-26

7.1 Rectangle - Laplace

Q: Can we use separation of variables for $\nabla^2 u = 0$ with all nonhomogeneous BCs, and if so, under what conditions?

We can use separation of variables, however particular solutions that individually satisfy each nonhomogeneous BC must be added together with superposition.

Q: Why is superposition of particular solutions justified in the context of Laplace's Eq.?

Because Laplace's Eq. is a linear PDE ($\mathcal{L}(u) = 0$).

Q: Let $\phi'' + \lambda\phi = 0$ with $\phi = \phi(x)$, $x \in [0, L]$, and $\phi(0) = \phi(L) = 0$. What are the eigenvalues and eigenfunctions?

The ODE solution is $\phi(x) = c_0 \sin\left(\sqrt{\lambda}x\right) + c_1 \cos\left(\sqrt{\lambda}x\right)$. Plug in the BCs.

$$\lambda_n = \left(\frac{n\pi}{L}\right)^2, \quad n \in \mathbb{Z}^+$$
$$\phi_n(x) = \sin\left(\frac{n\pi}{L}x\right)$$

Derivation for Laplace's Eq. in a rectangle:

$u(x, y) = u_{f0} + u_{f1} + u_{g0} + u_{g1}$

$u_{g0}(x, 0) = g_0(x), \quad u_{g0}(x, H) = u_{g0}(0, y) = u_{g0}(L, y) = 0.$

Using superposition,

$$u(x, y) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi}{L}x\right) \sinh\left(\frac{n\pi}{L}(y - H)\right)$$

.

To find the coefficients, A_n, we use orthogonality.

$$g_0(x) := u(x, 0) = \sum_{n=1}^{\infty} A_n \sin\left(\frac{n\pi}{L}x\right) \sinh\left(\frac{n\pi}{L}(-H)\right)$$

$$\implies A_n = \frac{1}{\sinh\left(\frac{n\pi}{L}(-H)\right)} \frac{2}{L} \int_0^2 \sin\left(\frac{n\pi}{L}x\right) g_0(x) \mathrm{d}x$$

7.2 Circular Disk - Laplace

$u = u(r, \theta), r \in (0, R), \theta \in (-\pi, \pi)$. PDE:

$$\nabla^2 u(r, \theta) = 0 = \frac{1}{r}\partial_r(r\partial_r u) + \frac{1}{r^2}\partial_\theta^2 u = \frac{1}{r}\partial_r u + \partial_r^2 u + \frac{1}{r^2}\partial_\theta^2 u$$

BCs:

$$u(R,\theta) = \Theta(\theta) \qquad \text{(Outer edge B.C.)}$$
$$|u(0,\theta)| < \infty \qquad \text{(finite at origin)}$$
$$u(r,\pi) = u(r,-\pi) \qquad \text{(periodic I)}$$
$$\partial_\theta u(r,\pi) = \partial_\theta u(r,-\pi) \qquad \text{(periodic II)}$$

Separate variables:

$$u(r,\theta) = G(r)\Theta(\theta) \quad \nabla^2 u = 0$$
$$0 = \Theta\frac{1}{r}\partial_r(r\partial_r G) + \frac{1}{r^2}G\partial_\theta^2\Theta$$
$$0 = \frac{1}{Gr}\partial_r(r\partial_r G) + \frac{1}{r^2\Theta}\partial_\theta^2\Theta = -\lambda + \lambda$$
$$\therefore \boxed{\Theta'' + \lambda\Theta = 0}$$
$$\boxed{r\partial_r(r\partial_r G) - \lambda G = 0}$$

BCs in θ:

$$\partial_\theta^2\Theta + \lambda\Theta = 0.\ \ \Theta(\pi) = \Theta(-\pi).\ \ \Theta'(\pi) = \Theta'(-\pi)$$
$$\therefore \quad \boxed{\lambda_n = n^2, n \in \mathbb{N}.\ \ \phi_n = c_0\sin(n\theta) + c_1\cos(n\theta)}$$

BCs in r:

$$r\partial_r(r\partial_r G) - \lambda G = 0$$
$$\implies r^2G'' + rG' - \lambda G = 0$$

Let $G(r) = r^p$.

$$(p(p-1) + p - \lambda)r^p = 0$$
$$p^2 = \lambda = n^2. \implies p = \pm n$$
$$\therefore\ G(r) = r^{\pm n}$$

In order to figure whether to take + or - n, impose the finite boundary condition:

$$|G(0)| < \infty. \lim_{r\to 0} r^n =; 0. \quad \lim_{r\to 0} r^{-n} = \infty$$

Thus, $G(r) = r^n$.

So far, the general solution is

$$u(r,\theta) = G(r)\Theta(\theta) = \sum_{n=0}^{\infty} A_n r^n\cos(n\theta) + \sum_{n=1}^{\infty} B_n r^n\sin(n\theta)$$

Last BC at $r = R$:

$$f(\theta) := u(R, \theta) = \sum_{n=0}^{\infty} A_n R^n \cos(n\theta) + \sum_{n=1}^{\infty} B_n R^n \sin(n\theta)$$

$$A_0 = \frac{\int f(\theta) \mathrm{d}\theta}{\int \mathrm{d}\theta} = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(\theta) \mathrm{d}\theta$$

$$A_n = \frac{1}{\pi R^n} \int_{-\pi}^{\pi} f(\theta) \cos(n\theta) \mathrm{d}\theta$$

$$B_n = \frac{1}{\pi R^n} \int_{-\pi}^{\pi} \ldots$$

Chapter 8 Fourier Series

Book §3

The Fourier series of $f(x)$ on $x \in [-L, L]$ is

$$f \approx a_0 + \sum_{n=1}^{\infty} a_n \cos\left(\frac{n\pi}{L}x\right) + \sum_{n=1}^{\infty} b_n \sin\left(\frac{n\pi}{L}x\right)$$

$$a_0 = \frac{1}{2L}\int_{-L}^{L} f(x)\mathrm{d}x$$

$$a_n = \frac{1}{L}\int_{-L}^{L} f(x)\cos\left(\frac{n\pi}{L}x\right)\mathrm{d}x$$

$$b_n = \frac{1}{L}\int_{-L}^{L} f(x)\sin\left(\frac{n\pi}{L}x\right)\mathrm{d}x$$

Thm: f piecewise smooth $\implies$ f has finitely many corners and jumps. 2 results from this

Example 8.1

$$f(x) = \begin{cases} 0 & x \in [-\pi, 0) \\ 2 & x \in [0, \pi] \end{cases}$$

Fourier Coefficients

$$a_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(x)\mathrm{d}x$$

$$a_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\cos(nx)\mathrm{d}x = \frac{1}{\pi}\int_{0}^{\pi} 2\cos(nx)\mathrm{d}x$$

$$b_n = \frac{1}{\pi}\int_{-\pi}^{\pi} f(x)\sin(nx)\mathrm{d}x$$

$$\begin{aligned} f &\approx 1 + \sum_{n \text{ odd}} \frac{4}{n\pi}\sin(nx) \\ &= 1 + \sum_{k=0}^{\infty} \frac{4}{(2k+1)\pi}\sin(nx) && (\text{let } n = 2k+1) \end{aligned}$$

At $x = 0$, f is not continuous. We have $f(0) = 2$.

At $x = \frac{\pi}{2}$:

$$2 = 1 + \frac{4}{\pi}\sum_{n=1}^{\infty} \frac{1}{2k+1}\sin\Big((2k+1)\frac{\pi}{2}\Big)$$

$$\sum \frac{\sin\big((2k+1)\frac{\pi}{2}\big)}{2k+1} = \frac{\pi}{2}$$

$$\sin\left(\frac{(2k+1)\pi}{2}\right)...$$

8.1 Fourier Series of Odd/Even Functions

An odd function integrated over symmetric interval is 0. Thus, the Fourier series of an odd function only has $\sin()$ terms. Similarly, the Fourier series of an even function has only $\cos()$ terms.

Q: When does the Fourier Series have discontinuities?

General Fourier Series: A general Fourier Series for f on $x \in (-L, L)$ is continuous as long as $f(x)$ is continuous and $f(-L) = f(L)$.

Cosine Series: A coine series will be continuous as long as f is continuous and $f(L) = f(-L)$ with f extended in an even way

Sine Series: A sine series will be continuous if f is continuous, $f(0) = 0$, and $f(L) = 0$.

Term by term time differentiation

When can one term by term differentiate a Fourier Series w.r.t. x?

8.2 Fourier Sine Series

Recall that the temperature $u(x,t)$, in a 1-D rod $x \in (0, L)$ with $u(0,t) = u(L,t) = 0$ satisfies

$$u(x,t) = \sum_{n=1}^{\infty} B_n \phi_n(x) e^{-\lambda_n k t}, \quad \sqrt{\lambda_n} = \frac{n\pi}{L}, \quad \phi_n(x) = \sin\Big(\sqrt{\lambda_n}x\Big).$$

The initial condition, $f(x) = \sum_n B_n \phi_n(x)$ is a series of sines, however our Fourier series defintion is defined over $x \in [-L, L]$, not $x \in [0, L]$. Also, $f(x)$ is not necessarily odd. In this sutation, we get the Fourier sine series by **extending** $f(x)$. The odd extension of $f(x)$ is piecewise smooth as long as f is piecewise smooth for $x \in [0, L]$.

Q: What is the Fourier sine series?

The Fourier sine series of $f(x)$ is the Fourier series of the odd extension of $f(x)$.

Q: $f(x) = x$ on $x \in [0, L]$. Derive the Fourier sine series of $f(x)$.

Sine series representation: $x \sim \sum_{n=1}^{\infty} B_n \sin\big(\frac{n\pi}{L}x\big), x \in [0, L]$. The sine series is equal to f on $x \in (-L, L)$ becuase there's no jump discontinuity at 0. If there was one, we could only say that $x = \sum_{n=1}^{\infty} B_n \sin\big(\frac{n\pi}{L}x\big), x \in (0, L)$.

Thus, the Fourier sine series of $f(x) = x$ is

$$x = \sum_{n=1}^{\infty} B_n \sin\Big(\frac{n\pi}{L}x\Big), x \in (-L, L).$$

$$B_n = \frac{1}{(\frac{L}{2})} \int_0^L f(x) \sin\Big(\frac{n\pi}{L}x\Big) \mathrm{d}x = \frac{2}{L} \int_0^L x \sin\Big(\frac{n\pi}{L}x\Big) = \frac{2L}{n\pi}(-1)^{n+1}.$$

Use integration by parts to do the integral.

Q: Given $f(x) = \cos\big(\frac{\pi}{L}x\big)$ for $x \in [0, L]$. What is the Fourier sine series of $f(x)$? Only set up the problem. You need not solve.

Fourier sine series representation: $\cos\big(\frac{\pi}{L}x\big) \sim \sum_{n=1}^{\infty} B_n \sin\big(\frac{n\pi}{L}x\big), x \in [0, L]$.

This function has jump discontinuities at both 0 and L, so the equality between f and its sine series holds only for $x \in (0, L)$ but not at $x = 0$ or $x = L$.

$$\therefore \quad \boxed{\cos\Big(\frac{\pi}{L}x\Big) = \sum_{n=1}^{\infty} B_n \sin\Big(\frac{n\pi}{L}x\Big), \quad x \in (0, L)}$$

$$\boxed{B_n = \frac{1}{(\frac{L}{2})} \int_0^L \cos\Big(\frac{\pi}{L}x\Big) \sin\Big(\frac{n\pi}{L}x\Big) \mathrm{d}x} = \begin{cases} 0 & n \text{ odd} \\ \frac{4n}{\pi(n^2-1)} & n \text{ even} \end{cases}$$

8.3 Appendix A: Orthogonality relations for sine and cosine

Asymmetric Boundaries

$$\int_0^L \sin\Big(\frac{n\pi}{L}x\Big) \sin\Big(\frac{m\pi}{L}x\Big) \mathrm{d}x = \frac{L}{2}\delta_{mn} = \begin{cases} 0 & m \neq n \\ L/2 & m = n, \end{cases}$$

$$\int_0^L \cos\Big(\frac{n\pi}{L}x\Big) \cos\Big(\frac{n\pi}{L}x\Big) \mathrm{d}x = \frac{L}{2}(1 + \delta_{n0})\delta_{mn} = \begin{cases} 0 & n \neq m \\ L/2 & n = m \neq 0 \\ L & n = m = 0 \end{cases}$$

Symmetric Boundaries

$$\int_{-L}^{L} \sin\left(\frac{n\pi}{L}x\right) \sin\left(\frac{m\pi}{L}x\right) \mathrm{d}x = L\delta_{nm} = \begin{cases} 0 & m \neq n \\ L & m = n \neq 0, \end{cases}$$

$$\int_{-L}^{L} \cos\left(\frac{n\pi}{L}x\right) \cos\left(\frac{n\pi}{L}x\right) \mathrm{d}x = L(1+\delta_{n0})\delta_{nm} = \begin{cases} 0 & n \neq m \\ L & n = m \neq 0 \\ 2L & n = m = 0 \end{cases}$$

$$\int_{-L}^{L} \sin\left(\frac{n\pi}{L}x\right) \cos\left(\frac{m\pi}{L}x\right) \mathrm{d}x = 0$$

Chapter 9 Sturm-Louivile

9.1 Exam 1 - §1-5

9.1.1 Exam 1 Info

- Today and Thursday not on exam
- Next Tuesday there's a review session that will basically be an office hours.
- There's a study guide on Courseworks. Ideally, you'll have done this before next Tuesday.
- Exam is next Thursday (Feb 25). No lecture. Exam is 90 minutes long and available from 9am EST Thursday to 9am EST on Friday. No typed solutions.
- There are free points available for completing an easy Gradescope "quiz". Must be done before 10am Monday Feb 22.

9.2 Higher-Dimensioal PDEs

Lecture 11, §7 of Haberman - post midterm

Wei Chung (TA) office hours now Thursday 4pm

7.4 Statements of Theorems for Helmholtz Eq.

1. There may be eigenfunctions ϕ_λ for a single λ that are lineraly independent.
2. If $\lambda_1 \neq \lambda_2$, then ϕ_{λ_1} and ϕ_{λ_2} are orthogonal.

$$\nabla^2\phi + \lambda\phi = 0 \text{ in region } R$$

$$\text{on bounary } \delta R$$

§7.5 Haberman

Part III

Data Mining

Chapter 10 Attention with Performers [Lec. 2]

We're going to talk about exciting applications with sequential data. Today, we'll focus on bioinformatics. Last time we talked about transformers. Today, we'll

softmax attention:

- Comprehensive Guide to the Attention Mechanism [blog post]
- Attention is all you need [paper].

Why do we need better memorization and attention in ML?

- "Developmental Robotics: A Complex Dynamical SYstem ith Several Spatiotemporal scales."
- Memory is key to AI and currently existing sequential RNNs fail to memorize well.
- Attention dimensions: spatial and temporal. - read the paper.
- Standard attention mechanisms are effectively parallelizable and avoid catastrophic forgetting but are not scalable. "It used more memory and more computation per real interaction..." - DeepMind nav by sight.
- 2 applications: 1. DeepMind policy nagivating simply by sight. 2. Robotic arm solving Hanoi towers.

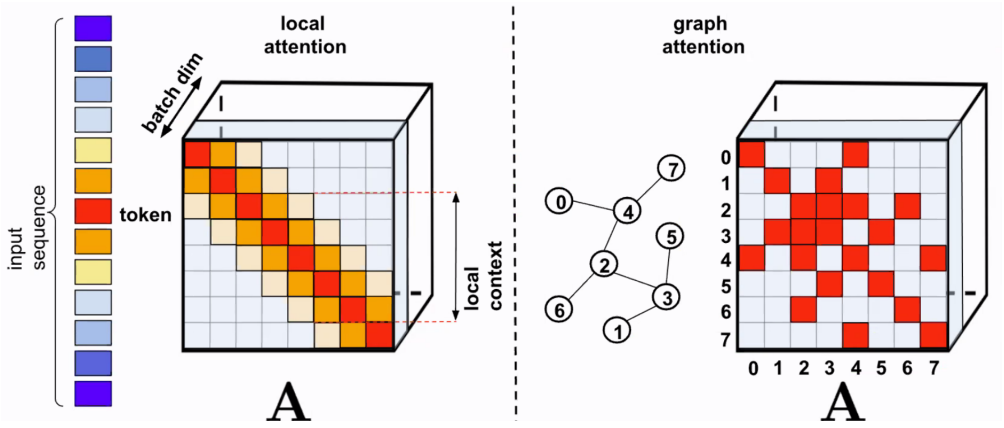
What attention mechanisms essentially do is take a sequence of tokens and learn the relationships between them. We say that the model learns how a token "attends" another token.

Attention matrix: Think of the attention matrix as a similarity matrix between tokens. The elements are scalars that capture the relevance between tokens.

A sequence consists of L tokens. Each token has dimension d. The attention matrix, A, is $A \in M_{L\times L}$. We perform transformtions on the input data seq (which is L by d) by multiplying either by $W_Q \in \mathcal{T}_{D_B\times d\times L}$, where D_B is the batch dimension and $\mathcal{T}$ denotes the space of tensors (I'm thinking of them as Tensorflow tensors or PyTorch tensors). Then,

$$\text{softmax attention } = \exp\left(\frac{\boldsymbol{q}\boldsymbol{k}^T}{d_{QK}}\right)$$

QK-pairs refer to "queries" and "keys". Space and time complexity is quadratic in the number of tokens, i.e. $O(L^2)$. So, it's not super scalable.



One way to make the attention architecture more scalable would be to only look at local attention, or attention in the neighborhood of each token. Graph attention is another possibility.

We talked about how the attention matrix could roughly be though of as a similarity matrix. In ML, we refer to similarity matrices as kernels.

Terms to look up:

- dense attention vs. sparse attention
- attend (verb)
- attention matrix
- queries and keys in attention
- performer model
- kernelizable in "attention is kernelizable". Also, kernel methods in general
- partition function

Usually, you take some row representation of you data.

Chapter 11 Transformers (cont.) [Lec. 5]

Feb. 22

Positional encoding

Attention Is All You Need. This section refers to positional encoding You have a positional encoding, add it to your one hot vector. That's how you encode time. That's how you encode positions in your sequence.

Enriched row embedding

Gieven x_t, a one-hot vector, and $f(t)$, where t is a positional encoding. We add $x_t + f(t)$.

Multiple-head mechanism

- Multi-Head Attention [article] - Lilian Weng
- [PyTorch code]

Multi-head Attention is a module for attention mechanisms which runs through an attention mechanism several times in parallel. The independent attention outputs are then concatenated and linearly transformed into the expected dimension. Intuitively, multiple attention heads allows for attending to parts of the sequence differently (e.g. longer-term dependencies versus shorter-term dependencies).

$$\text{MultiHead}(Q, K, V) = HW_0 \tag{11.1}$$

$$H = [h_1, \ldots, h_h], \quad h_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{11.2}$$

The above W terms are learnable parameter matrices and h_i denotes the ith "head". Note that scaled dot-product attention is most commonly used in this module, although in principle, it can be swapped out for other types of attention mechanism.

The multi-head attention mechanism was introduced by Vaswani et al. in Attention Is All You Need.

L is the length of the sequence, b is the batch dim, and d is the dimensionality of the token embedding. The data is an $L \times b \times d$ tensor.

As in standard NNs, transformers process data in batches.

The idea of multiple-head attention is to use a couple different attention models in parellel.

If you take a specific attention model, what you're essentially doing is learning 3 matrices, W_Q, W_X, and W_V. Let $\mathcal{T}$ be the space of tensors. $W_Q \in \mathcal{T}_{L \times d_{QK} \times b}, W_K \in \mathcal{T}_{L \times d_{QK} \times b}, W_V \in \mathcal{T}_{L \times d \times b}$.

Terms to look up:

- positional encoding

- mutliple-head mechanism

Chapter 12 The Unreasonable Effectiveness of ES

A Tale of Hadamard-Minitaurs and Toeplitz-Walkers.

Fancy finite difference replacing backpropagation