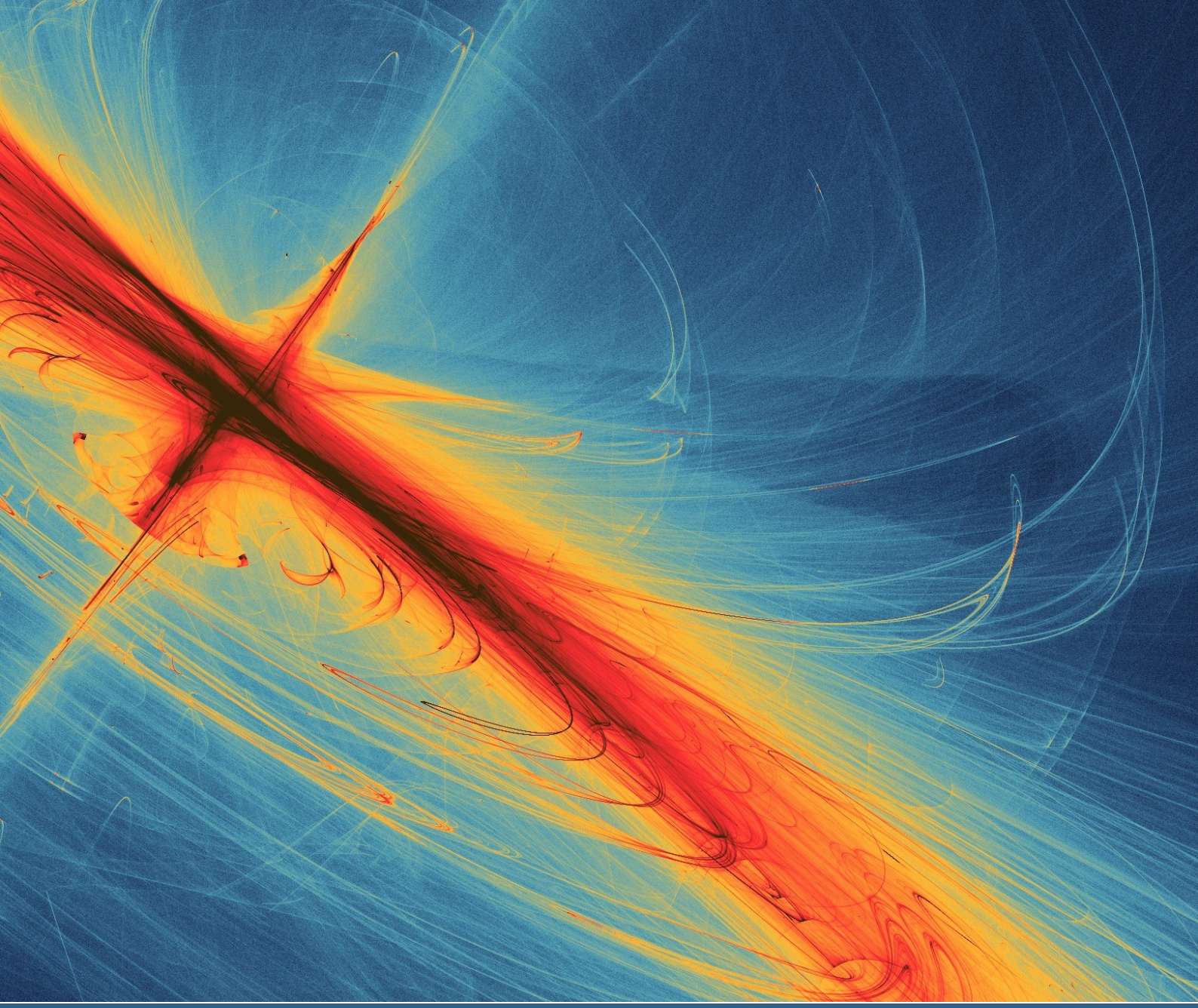


Notes: Spring 2021

Data Science, Stochastic Analysis, & PDEs

Author: Unique Divine

Institute: Columbia University

Date: Spring 2021

Contents

Part I

Stochastic Analysis

Chapter 1 Introduction, Intelligent Agents

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

1.1.1 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.

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

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

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

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

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

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

1.2 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

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

$$\implies (2b\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2] \cdot b^2\mathbb{E}[Y^2] \leq 0$$

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

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.

1.2.1 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.

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

1.3.1 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.

1.3.2 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 $\implies X_n \xrightarrow{D} X$.
- $X_n \xrightarrow{\ell^p} X \implies X_n \xrightarrow{\mathbb{P}} X$.
- $X_n \xrightarrow{\ell^p} X \implies X_n \xrightarrow{\ell^q} X, \quad 1 \leq q \leq p$.

1.4 Markov Chains [Lec. 3, Jan 25]

Stochastic process:

$X_t(\omega), t \in T$. 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.

This is a very wide set of functions, so we'll restrict our focus to something more specific to build intuition.

1.4.1 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.

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

Part II

PDE

Chapter 2 ODE Solving

2.1 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.

2.1.1 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.

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

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

2.1.2 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 3 Heat Equation

3.1 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 ($n \geq 1$).

$$u(x,0) = f(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos\Big(\frac{n\pi}{L}x\Big) \tag{3.1}$$

Q: Solve for A_0.

Integrate and interchange the order of sum and integral.

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

$$A_0 = \frac{1}{L}\int_0^L f(x)\mathrm{d}x \tag{3.2}$$

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:

$$\therefore \quad A_0 = \frac{1}{L}\int_0^L f(x)\mathrm{d}x. \tag{3.3}$$

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

BCs: Periodic boundary conditions

Chapter 4 Laplace's Eq.

- Book §2.5
- Start: Lecture 4, 1-26

4.1 Rectangle

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 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: What are dirichlet BCs in the context of $\nabla^2 u(x, y) = 0$?

$$u(0, y) = f_0(y)$$
$$u(L, y) = f_1(y)$$
$$u(x, 0) = g_0(x)$$
$$u(x, H) = g_1(x)$$

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.

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

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

4.2 Circular Disk

PDE:

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

$$r \in (0, R)$$

BCs:

$$u(R,\theta) = \Theta(\theta) \qquad \text{(Outer edge depends only on angle)}$$

$$|u(0,\theta)| < \infty \qquad \text{(finite at origin)}$$

$$u(r,\pi) = u(r,-\pi) \qquad \text{(periodic I)}$$

$$\partial_\theta u(r,\pi) = \partial_\theta u(r,-\pi) \qquad \text{(periodic II)}$$

Separate variables:

$$u(r,\theta) = G(r)\Theta(\theta) \quad \nabla^2 u = 0$$

$$0 = \Theta\frac{1}{r}\partial_r(r\partial_r G) + \frac{1}{r^2}G\partial_\theta^2\Theta$$

$$0 = \frac{1}{Gr}\partial_r(r\partial_r G) + \frac{1}{r^2\Theta}\partial_\theta^2\Theta = -\lambda + \lambda$$

$$\therefore \boxed{\Theta'' + \lambda\Theta = 0}$$

$$\boxed{r\partial_r(r\partial_r G) - \lambda G = 0}$$

Evaluate with BCs in θ:

$$\partial_\theta^2\Theta + \lambda\Theta = 0.\ \ \Theta(\pi) = \Theta(-\pi).\ \ \Theta'(\pi) = \Theta'(-\pi)$$

$$\therefore \quad \boxed{\lambda_n = n^2, n \in \mathbb{N}.\ \ \phi_n = c_0\sin(n\theta) + c_1\cos(n\theta)}$$

Evaluate with BCs in r:

$$r\partial_r(r\partial_r G) - \lambda G = 0$$

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

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 $r = R$:

$$f(\theta) := u(R,\theta) = \sum_{n=0}^{\infty} A_n R^n \cos(n\theta) + \sum_{n=1}^{\infty} B_n R^n \sin(n\theta)$$
$$A_0 = \frac{1}{2\pi}\int_{-\pi}^{\pi} f(\theta)\mathrm{d}\theta$$
$$A_n = \frac{1}{\pi R^n}\int_{-\pi}^{\pi} f(\theta\cos(n\theta))\mathrm{d}\theta$$
$$B_n = \frac{1}{\pi R^n}\int_{-\pi}^{\pi} \ldots$$

Part III

Data Mining

Chapter 5 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.

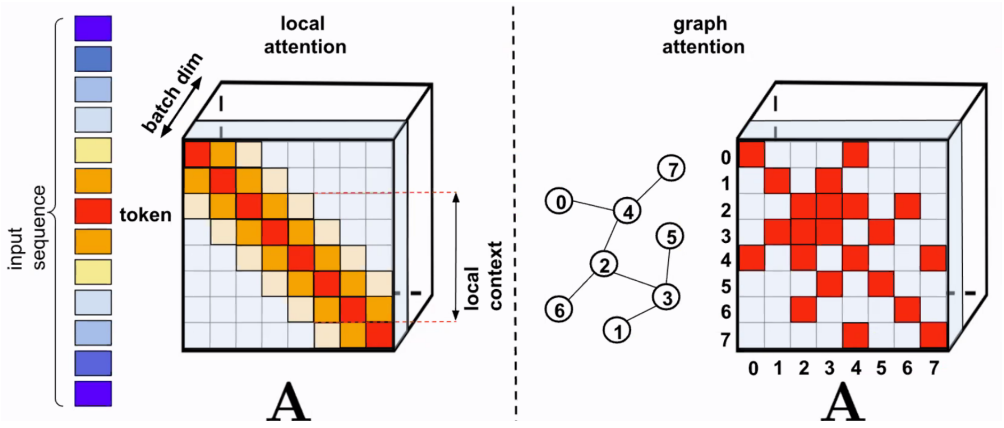
What attention mechanisms essentially do is take a sequence of tokens and learn the relationships between them. We say that the model learns how a token "attends" another token.

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.