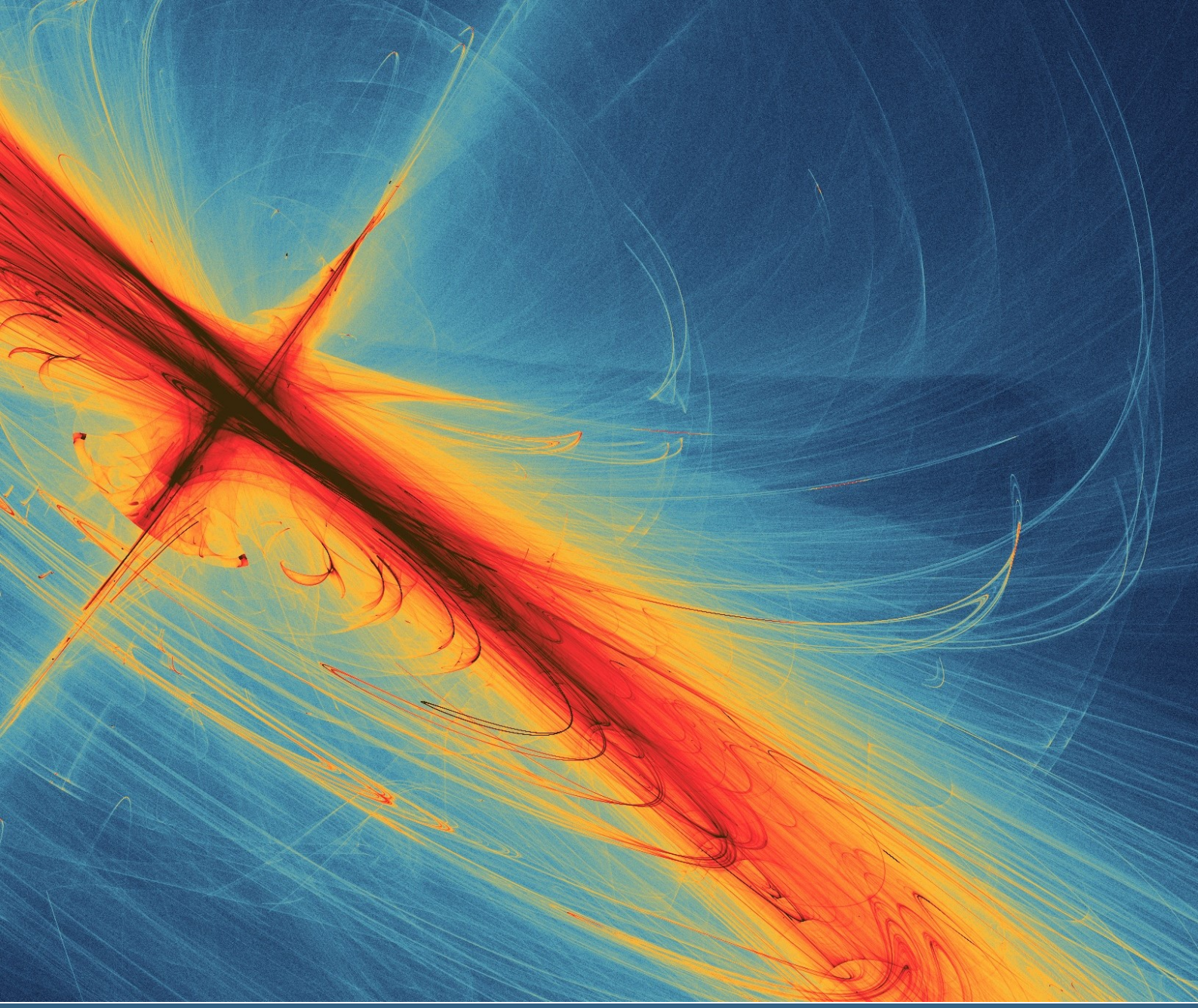

Notes: Spring 2021

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

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

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

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

Part III

Data Mining

Chapter 4 Introduction, Intelligent Agents

4.1 Lec. 2

If a function has one ontinuous derivative, then f is convex over a convex set S.

$$\iff f(y) \geq f(x) + \nabla f(x)^T (y - x), \forall x, y \in S. \tag{4.1}$$

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n} \right) \tag{4.2}$$

$$\tag{4.3}$$

Proof "⇒":

Since f is convex on S, then $\forall x, y \in S$ and $0 < \alpha \leq 1$,

$$f(\alpha y + (1-\alpha)x) \leq \alpha f(y) + (1-\alpha) f(x) \tag{4.4}$$

$$\implies \frac{f(\alpha y + (1-\alpha)x)}{\alpha} \leq f(y) + \frac{(1-\alpha) f(x)}{\alpha} \tag{4.5}$$

$$\implies \frac{f(x + \alpha(y-x))}{\alpha} - \frac{f(x)}{\alpha} \leq f(y) - f(x) \tag{4.6}$$

$$\implies \boxed{\frac{f(x + \alpha(y-x)) - f(x)}{\alpha}} \leq f(y) - f(x) \tag{4.7}$$

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} = \nabla f(x) \cdot (y - x) = \nabla f(x)^T (y - x) \tag{4.8}$$

$$\implies f(y) \geq f(x) + \nabla f(x)^T (y - x) \tag{4.9}$$

$$\tag{4.10}$$

"⇐":

Let $t = \alpha x + (1-\alpha) y, \quad 0 \leq \alpha \leq 1$. Then, $t \in S$ (S is convex and $x, y \in S$).

$$\implies f(y) \geq f(t) + \nabla f(t)^T (y - t)$$

$$f(x) \geq f(t) + \nabla f(t)^T (x - t)$$

$$\therefore \quad \alpha f(x) + (1-\alpha) f(y) \geq [\alpha + (1-\alpha)] f(t) + \alpha \nabla f(t)^T (x - t) + (1-\alpha) \nabla f(t)^T (y - t) = f(t) + \alpha \nabla f(t)^T x$$

$$= f(t) = f(\alpha x + (1-\alpha) y)$$

$\implies$ f is a convex fn on a convex set S.

Case II: If a one-dim fn f is a twice differentiable (two continuous derivatives) then f is convex on a convex set S, i.e.

$$f''(x) \geq 0, \quad \forall x \in S.$$

In a multidim case, we define a hessian matrix.

$$\nabla^2 f(x) = \begin{pmatrix} f_{x_1x_1} & \cdots & f_{x_1x_n} \\ \vdots & \ddots & \vdots \\ f_{x_nx_1} & \cdots & f_{x_nx_n} \end{pmatrix}$$

If $y^T \nabla^2 f(x) y \geq 0 \; \forall y \neq 0$ (Hessian matrix is positive semi-definite), then ($\iff$) f is convex on a convex set S. Alternatively, we can check the eigenvalues of $\nabla^2 f(x)$

Convex optimization problem: Recall that the $\max f(x) = \min -f(x)$. The convex optimization problem is defined by the following scenario. We hope to

$$\text{Find } \min f(x) \text{ s.t.}$$

$$g_i(x) \leq 0, i \in I$$

$$g_i(x) = 0, i \in \epsilon,$$

where $f(x)$ is convex, $g_{i \in I}$ are convex, and $g_{i \in \epsilon}$ are affine.

First, let's show that $D_1 = \{x | g_{i \in I}(x) \leq 0\}$ is a convex set. This means that $\forall x_1, x_2 \in D_1$, we want to verify

$$\alpha x_1 + (1-\alpha) x_1 \in D_1, \; 0 \leq \alpha \leq 1,$$

$$g_i(\alpha x_1 + (1-\alpha) x_2) \leq \alpha g_i(x_1) + (1-\alpha) g_i(x_2) \text{ and } g_i(x_1), g_i(x_2) \leq 0$$

Second, for the affine function $f(x) = a^T x + b, \; a \in \mathbb{R}^n, b \in \mathbb{R}$,

$$D_2 = \{x | g_{i \in \epsilon}(x) = 0\}$$

is a convex set.

$$\begin{aligned} g_i ... &= \alpha(a^T x_1 + b) + (1-\alpha)(a^T x_2 + b) \\ &= \alpha g_i(x_1) + (1-\alpha) g_i(x_2) = 0 \\ &\implies \alpha x_1 + (1-\alpha) x_2 \in D_2. \end{aligned}$$

If D_1 is convex and D_2 is convex, then $D_1 \cap D_2$ is convex (exercise).

$\implies$ S is convex.

THm:

Global solution of convex optimization problems. Let x_* be a local minimizer of a convex optim problem. Then x_* is also a global minimizer. If the objective function is strictly convex, then x_* is the unique global minimizer.

Proof (By contradiction) Let x_* be a local minimizer and suppose it is not a global minimizer. Then, there exists some point $y \in S$ s.t. $f(y) < f(x_*)$.

Take $0 < \alpha < 1$, then f is convex on S.

$$\begin{aligned} f(\alpha x_* + (1-\alpha)y) &\leq \alpha f(x_*) + (1-\alpha)f(y) \\ &< \alpha f(x_*) + (1-\alpha)f(x_*) \\ &= f(x_*). \end{aligned}$$

This means there are points $\alpha x_* + (1-\alpha)y$ that are arbitrarily close to x_* ($\alpha \to 1$) s.t. $f(\alpha x_* + (1-\alpha)y) < f(x_*)$ (contradiction with local minimizer).

Proof for global minimizer (exercise) - hint: use contradiction.

General optimization algorithm (iterative methods).

Algorithm 1:

1. Input the initial guess x_0.
2. For $k = 0, 1, \ldots,$
   (a). If x_k is optimal, stop. (test optimality)
   (b). Determine x_{k+1}. Update $x_k \to x_{k+1}$. (determine new points)

Algorithm 2:

1. Input initial guess x_0.
2. For $k = 0, 1, \ldots,$
   (a). If x_k is optimal, stop.
   (b). Determine a search direction, p_k
   (c). Determine a step length α_k that leads to an improved estimation of the solution: $x_{k+1} = x_k + \alpha_k p_k$.

In the above algorithm, p_k is called the **descent direction** and α_k the **line search**.

- For an unconstrained optimization problem, we typically require p_k to be a descent direction of the function of f at point x_k s.t.

$$f(x_k + \alpha p_k) < f(x_k),\ 0 < \alpha < \epsilon.$$

- For a constrained problem,

$$f(x_k + \alpha p_k) < f(x_k) \text{ and } x_k + \alpha p_k \in S,\ \alpha \in [0, \epsilon],$$

  where ϵ is a small positive number.
- After we have p_k, then $\min_{\alpha \geq 0} f(x_k + \alpha p_k)$.

When we have a convergent algorithm and want to quantify how fast it converges, we describe this with **rate of convergence**. The rate of convergence describes how quickly the estimates of the solution approach the exact solution.

We say that the sequence x_k converges to x_* with rate $r \geq 1$ and rate constant c if

$$\lim_{k \to \infty} \frac{\|e_{k+1}\|}{\|e_k\|^r} = c \text{ and } c < \infty.$$

$$e_k = x_k - x_*, \quad e_{k+1} = x_{k+1} - x_* \qquad (\lim_{k\to\infty} x_k = x_*)$$

$$\|e_{k+1}\| \approx c\|e_k\|^r, \quad \|e_k\| \approx c\|e_{k-1}\|^r$$

$$\implies \quad \frac{\|e_{k+1}\|}{\|e_k\|} \approx \left(\frac{\|e_k\|}{\|e_{k-1}\|}\right)^r$$

$$\implies \quad r_k \approx \frac{\log \frac{\|e_{k+1}\|}{\|e_k\|}}{\log \frac{\|e_k\|}{\|e_{k-1}\|}}$$

$r = 1$ is linear convergence

- $0 < c < 1$ is error reduced by a constant factor.
- $c > 1$ is divergence.
- $c = 1$ is oscillating
- $c = 0$ is superlinear convergence

$r = 2$ quadratic convergence