



# **Notes: Applied Stochastic Analysis (APMA 4990)**

**Taught by Kui Ren**

**Author:** Unique Divine

**Institute:** Columbia University

**Date:** Spring 2021



# Contents

<b>I</b>	<b>Stochastic Analysis</b>	<b>1</b>
<b>1</b>	<b>Introduction, Intelligent Agents</b>	<b>2</b>
1.1	Probability Theory I [Lec. 0, Jan 11] . . . . .	2
1.2	Probability Theory II [Lec. 1, Jan 13] . . . . .	4
1.3	[Lec. 2, Jan 20] . . . . .	6
<b>II</b>	<b>PDE</b>	<b>7</b>
<b>2</b>	<b>Heat Equation</b>	<b>8</b>
2.1	Lec. 1: Syllabus & Logistics . . . . .	8
2.2	Equilibrium Temperature distribution (Haberman §1.4) . . . . .	8
<b>III</b>	<b>Data Mining</b>	<b>10</b>
<b>3</b>	<b>Introduction, Intelligent Agents</b>	<b>11</b>
3.1	Lec. 2 . . . . .	11

## **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,  $\sigma$ -algebra, probability measure

A sigma algebra 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.  $\mathbb{P} : f \rightarrow [0, 1]$ .

- $E_1 \subseteq E_2 \implies \mathbb{P}(E_1) \leq \mathbb{P}(E_2)$
- Boole's Inequality:  $\mathbb{P}(\bigcup_{i=1}^{\infty} E_i) \leq \sum_i \mathbb{P}(E_i)$
- 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,  $X(\omega) : \Omega \rightarrow \mathbb{R}$ .

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

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

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

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

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

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

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:  $\phi$  is monotone increasing

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

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

$$\begin{aligned} &\implies Y = |x - \mathbb{E}[x]| \\ &\implies \mathbb{P}(|x - \mathbb{E}[x]| \geq a) \leq \frac{\mathbb{E}(|x - \mathbb{E}[x]|^2)}{a^2} \end{aligned}$$

Why is this useful? It means that if you know how to control the variance, then you know how to control the probability. In the more general case,  $\phi$  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

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

## 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  $(\Omega, \mathcal{F}, \mathbb{P})$ . 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,  $M_X(t) \equiv \mathbb{E}[e^{tX}]$ ,  $t \in \mathbb{R}$ .

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

Example 1.

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

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

Example 2.

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

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

**Thm:**

$$\phi(0) = 1 \tag{1.1}$$

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

**Proof**

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

**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.

$\phi_z(t) = \phi_{x_1} + \dots + \phi_{x_n}$ . 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,  $\forall k \leq n$ ,  $\phi^{(k)}(t) = i^k \int x^k e^{itx} dF(x)$

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

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

### 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  $\{x_n\}_{n \in \mathbb{N}}$  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  $n \rightarrow \infty$ .

definition of "convergence in probability"

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

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

### Kinchtin Weak LLN (Thm):

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



**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 [Lec. 2, Jan 20]



**Part II**

**PDE**

## Chapter 2 Heat Equation

### 2.1 Lec. 1: Syllabus & Logistics

Prof will go over the expectations: expectations in terms of time commitment, workload, etc. We'll also talk about what you'll get out of this course.

### 2.2 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  $0 \leq x \leq L$  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 equilibrium 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  $\partial_x^2 u = 0$ .**

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

**Q: For equilibrium 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:**

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

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

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

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

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

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

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

**Q:**

## **Part III**

# **Data Mining**

## Chapter 3 Introduction, Intelligent Agents

### 3.1 Lec. 2

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

$$\iff f(y) \geq f(x) + \nabla f(x)^T(y - x), \forall x, y \in S. \quad (3.1)$$

$$\nabla f(x) = \left( \frac{\partial f}{\partial x_1}, \dots, \frac{\partial f}{\partial x_n} \right) \quad (3.2)$$

$$(3.3)$$

**Proof** " $\Rightarrow$ ":

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

$$f(\alpha y + (1 - \alpha)x) \leq \alpha f(y) + (1 - \alpha)f(x) \quad (3.4)$$

$$\implies \frac{f(\alpha y + (1 - \alpha)x)}{\alpha} \leq f(y) + \frac{(1 - \alpha)f(x)}{\alpha} \quad (3.5)$$

$$\implies \frac{f(x + \alpha(y - x))}{\alpha} - \frac{f(x)}{\alpha} \leq f(y) - f(x) \quad (3.6)$$

$$\implies \boxed{\frac{f(x + \alpha(y - x)) - f(x)}{\alpha}} \leq f(y) - f(x) \quad (3.7)$$

$$\lim_{\alpha \rightarrow 0} \frac{f(x + \alpha(y - x)) - f(x)}{\alpha} = \nabla f(x) \cdot (y - x) = \nabla f(x)^T(y - x) \quad (3.8)$$

$$\implies f(y) \geq f(x) + \nabla f(x)^T(y - x) \quad (3.9)$$

$$(3.10)$$

" $\Leftarrow$ ":

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

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

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

$$\begin{aligned} \therefore \alpha f(x) + (1 - \alpha)f(y) &\geq [\alpha + (1 - \alpha)]f(t) + \alpha \nabla f(t)^T(x - t) + (1 - \alpha) \nabla f(t)^T(y - t) = f(t) + \alpha \nabla f(t)^T x \\ &= f(t) = f(\alpha x + (1 - \alpha)y) \end{aligned}$$

$\implies f$  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.

$$\nabla^2 f(x) = \begin{pmatrix} f_{x_1 x_1} & \cdots & f_{x_1 x_n} \\ \vdots & \ddots & \vdots \\ f_{x_n x_1} & \cdots & f_{x_n x_n} \end{pmatrix}$$

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

Find  $\min f(x)$  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

$$\alpha x_1 + (1 - \alpha)x_2 \in D_1, 0 \leq \alpha \leq 1,$$

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

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.

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

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  $x_*$  ( $\alpha \rightarrow 1$ ) 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  $k = 0, 1, \dots$ ,
  - (a). If  $x_k$  is optimal, stop. (test optimality)
  - (b). Determine  $x_{k+1}$ . Update  $x_k \rightarrow x_{k+1}$ . (determine new points)

### Algorithm 2:

1. Input initial guess  $x_0$ .
2. For  $k = 0, 1, \dots$ ,
  - (a). If  $x_k$  is optimal, stop.
  - (b). Determine a search direction,  $p_k$
  - (c). Determine a step length  $\alpha_k$  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  $\alpha_k$  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.

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

- For a constrained problem,

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

where  $\epsilon$  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

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

$$e_k = x_k - x_*, \quad e_{k+1} = x_{k+1} - x_*$$

$$(\lim_{k \rightarrow \infty} x_k = x_*)$$

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

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

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

$r = 1$  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