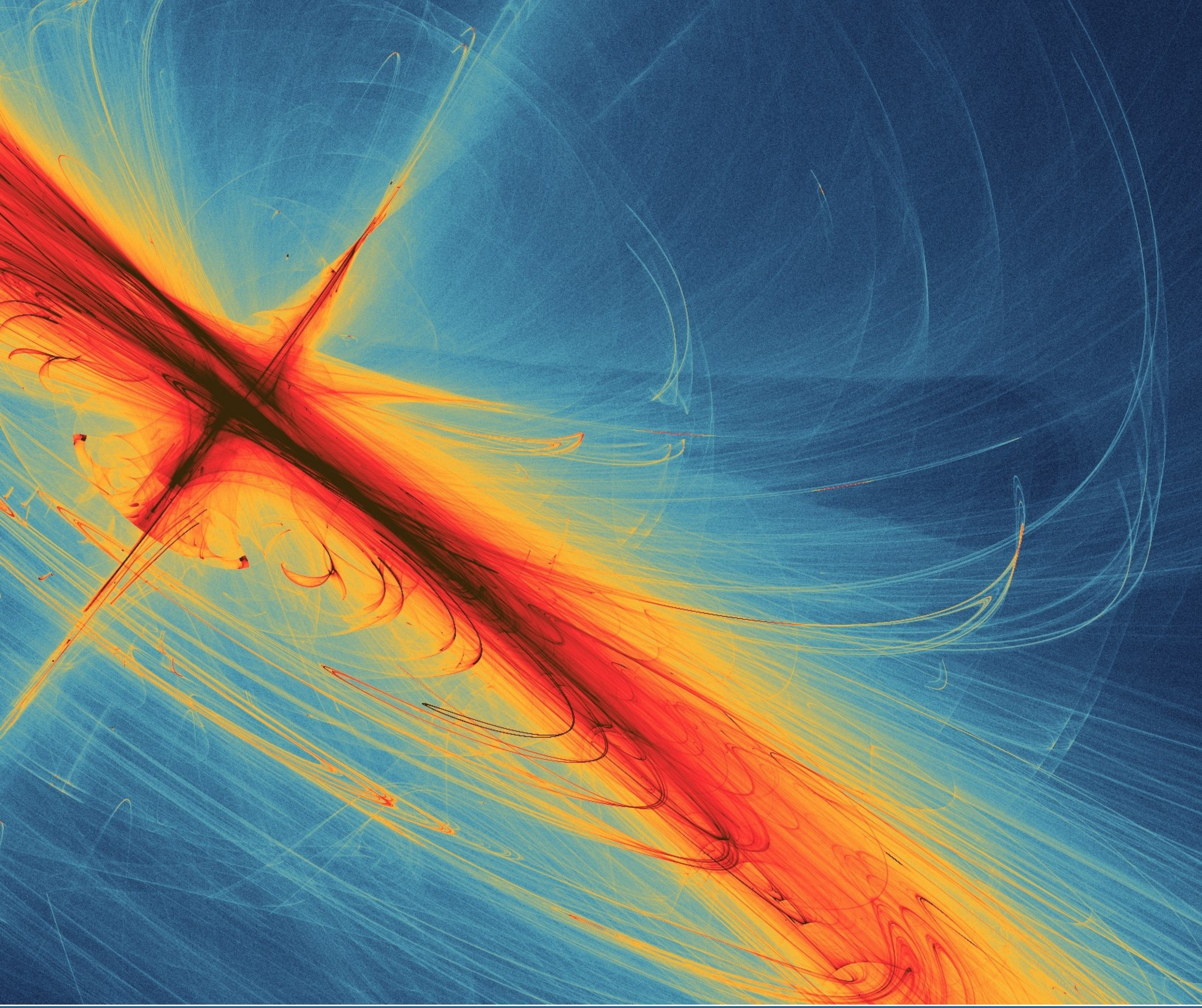

Notes: Applied Stochastic Analysis (APMA 4990)

Taught by Kui Ren

Author: Unique Divine

Institute: Columbia University

Date: Spring 2021

Contents

Part I

Stochastic Analysis

Chapter 1 Introduction, Intelligent Agents

1.1 Lec. 1: Syllabus & Logistics

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

Prof has taught this class twice before, so this is the third iteration.

Important shorthands I'll be using:

- Prof $\equiv$ the professor, Tony Dear
- h $\equiv$ hour(s); min $\equiv$ minute(s)

Plan for Today

- Syllabus and logistics
- Definition, foundations, and modern capabilities of AI
- Properties of **task environments**
- Structure and types of **intelligent agents**

Reading: Today's material will correspond to chapters 1 and 2 of the textbook.

Course Expectations

- MS-level (4k) CS course
  - Your peers: Mostly CS undergrads, CS grads, and SEAS grad
  - Some taking first 4k course, others first CS course
  - Must be able to learn independently, keep up with course reading
- **Coursework**:
  - Both **programming** AND **quantitative analysis (math: probability theory, maybe some linear algebra)**
- **Attendance**:
  - not required, but try your best to attend live
  - recordings uploaded by CVN within 24 hours of each lecture
- **Half-semester course / Immersive Course**:
  - covering **same material as full semester twice as fast** as usual
  - expect **workload equivalent to 2 regular courses**
  - University requires 18 h total weekly (6 h in class, 12 h outside of class) for immersive courses
- **Enrollment**
  - Section 001 (Fall A) closes Friday 9/11 and section 002 (Fall B) closes Friday 9/18

- **Auditing**: Students interested in auditing and who do not intend to enroll in Fall B may petition to audit by completing this form: https://forms.gle/x8xJugaNa4JcKE4G6. *If flexible, you're encouraged to audit Fall B –fewer students, better experience for everyone*

Grade Breakdown

Assessments: Quizzes (25%), Homework (40%), Final Exam (35%)

Grading: Grade determined based on two scales:

1. a fixed standard scale (A is 90%+, B is 80%+, etc.) as well as
2. a curved scale relative to all other students,

where the average of the latter is around a B to B+. We will calculate your grade according to both scales and give you the higher of the two. It is thus possible for everyone to receive an A in the course, but impossible for everyone to fail the course.

Next page includes:

- Course description
- Objectives
- Prerequisites
- Topics covered
- Reading Assignments
- Assessments and Grading
- Lectures and Quizzes
- Homework
- Late Submissions and Drops
- Regrading
- Academic Conduct and Integrity

1.2 Lec. 1: Jan 13

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.

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

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

Example 2.

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

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

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

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, $\dfrac{S_n}{n} \xrightarrow{\mathbb{P}} \mu := \mathbb{E}[x_n]$ 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 Reading: Artificial Intelligence Intro (Book ch.1)

Part II

Optimization

Chapter 2 Introduction, Intelligent Agents

2.1 Lec. 2

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

$$\iff f(y) \geq f(x) + \nabla f(x)^T(y-x), \forall x, y \in S. \tag{2.1}$$

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right) \tag{2.2}$$

$$\tag{2.3}$$

Proof "$\Rightarrow$":

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

$$f(\alpha y + (1-\alpha)x) \leq \alpha f(y) + (1-\alpha) f(x) \tag{2.4}$$

$$\implies \frac{f(\alpha y + (1-\alpha)x)}{\alpha} \leq f(y) + \frac{(1-\alpha)f(x)}{\alpha} \tag{2.5}$$

$$\implies \frac{f(x + \alpha(y-x))}{\alpha} - \frac{f(x)}{\alpha} \leq f(y) - f(x) \tag{2.6}$$

$$\implies \boxed{\frac{f(x + \alpha(y-x)) - f(x)}{\alpha}} \leq f(y) - f(x) \tag{2.7}$$

$$\lim_{\alpha \to 0} \frac{f(x + \alpha(y-x)) - f(x)}{\alpha} = \nabla f(x) \cdot (y-x) = \nabla f(x)^T(y-x) \tag{2.8}$$

$$\implies f(y) \geq f(x) + \nabla f(x)^T(y-x) \tag{2.9}$$

$$\tag{2.10}$$

"$\Leftarrow$":

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

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

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

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

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

$\implies$ 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.

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

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.

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

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

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

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

$$\begin{aligned}
&\|e_{k+1}\| \approx c\|e_k\|^r, \quad \|e_k\| \approx c\|e_{k-1}\|^r \\
&\implies \quad \frac{\|e_{k+1}\|}{\|e_k\|} \approx \left(\frac{\|e_k\|}{\|e_{k-1}\|}\right)^r \\
&\implies \quad r_k \approx \frac{\log \frac{\|e_{k+1}\|}{\|e_k\|}}{\log \frac{\|e_k\|}{\|e_{k-1}\|}}
\end{aligned}$$

$r = 1$ 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

Part III

Statistical Inference

Chapter 3 Introduction, Intelligent Agents

3.1 Lec. 1: Syllabus & Logistics

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

Prof has taught this class twice before, so this is the third iteration.

Important shorthands I'll be using:

- Prof $\equiv$ the professor, Tony Dear
- h $\equiv$ hour(s); min $\equiv$ minute(s)

Plan for Today

- Syllabus and logistics
- Definition, foundations, and modern capabilities of AI
- Properties of **task environments**
- Structure and types of **intelligent agents**

Reading: Today's material will correspond to chapters 1 and 2 of the textbook.

Course Expectations

- MS-level (4k) CS course
  - Your peers: Mostly CS undergrads, CS grads, and SEAS grad
  - Some taking first 4k course, others first CS course
  - Must be able to learn independently, keep up with course reading
- **Coursework**:
  - Both **programming** AND **quantitative analysis (math: probability theory, maybe some linear algebra)**
- **Attendance**:
  - not required, but try your best to attend live
  - recordings uploaded by CVN within 24 hours of each lecture
- **Half-semester course / Immersive Course**:
  - covering **same material as full semester twice as fast** as usual
  - expect **workload equivalent to 2 regular courses**
  - University requires 18 h total weekly (6 h in class, 12 h outside of class) for immersive courses
- **Enrollment**
  - Section 001 (Fall A) closes Friday 9/11 and section 002 (Fall B) closes Friday 9/18

- **Auditing**: Students interested in auditing and who do not intend to enroll in Fall B may petition to audit by completing this form: https://forms.gle/x8xJugaNa4JcKE4G6. *If flexible, you're encouraged to audit Fall B –fewer students, better experience for everyone*

Grade Breakdown

Assessments: Quizzes (25%), Homework (40%), Final Exam (35%)

Grading: Grade determined based on two scales:

1. a fixed standard scale (A is 90%+, B is 80%+, etc.) as well as
2. a curved scale relative to all other students,

where the average of the latter is around a B to B+. We will calculate your grade according to both scales and give you the higher of the two. It is thus possible for everyone to receive an A in the course, but impossible for everyone to fail the course.

Next page includes:

- Course description
- Objectives
- Prerequisites
- Topics covered
- Reading Assignments
- Assessments and Grading
- Lectures and Quizzes
- Homework
- Late Submissions and Drops
- Regrading
- Academic Conduct and Integrity

3.2 Lec. 1: Jan 13

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.

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

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

Example 2.

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

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

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

THm:

$$\phi(0) = 1 \tag{3.1}$$

$$|\phi(t)| \leq 1, \forall t \in \mathbb{R} \tag{3.2}$$

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.

3.2.1 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.

3.3 Reading: Artificial Intelligence Intro (Book ch.1)

Bibliography