MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Fall 2005 6.434J/15.085J

This document attempts to list the material covered in the class, but it might have a few omissions. It is meant to be a listing, not an exposition, and it is presented in rather abstract terms. For the final exam, you should understand the concepts, be able to manipulate them rigorously in proofs, and apply them to more practical problem of applied probability.

A reading list is available on the web for the material covered in class. You should consider reviewing your lecture notes, your problem sets with the solutions, the recitations and last year's final.

1 Set and measure theory

- Set theory. Correspondence between logical statement and set operations ($\forall \mapsto \bigcap, \exists \mapsto \bigcup$). For collection of sets A_i, $x \in \limsup A_i = \bigcap_{k=1}^{\infty} \bigcup_{i \geq k} A_i$ ($x \in A_n$ infinitely often) and $x \in \liminf A_i = \bigcup_{k=1}^{\infty} \bigcap_{i \geq k} A_i$ ($x \in A_n$ for all n except finitely many).

- Given a sample space Ω, σ-field $\mathcal{F}$ axioms and examples (Infinite coin tosses, Borel). σ-field generated by a collection of sets (e.g Borel $\mathcal{B}$ on $\mathbb{R}$ generated by $\{(-\infty, b),\ b \in \mathbb{R}\}$). Prove properties via a generator of a σ-field.

- Measure space $(\Omega, \mathcal{F}, \mu)$, e.g. $(\mathbb{R}, \mathcal{B}, \lambda)$ Lebesgue measure. Elements of $\mathcal{F}$ are measurable sets or events. Countable additivity (if A_i disjoint sets, then $\mu(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mu(A_i)$), continuity of measure/Fatou lemma (if A_i increasing sequence for inclusion, $\lim \mu(A_i) = \mu(\bigcup_{i=1}^{\infty} A_i)$ or for arbitrary sets A_i, $\limsup \mu(A_i) \leq \mu(\limsup A_i)$), union bound ($\mu(\bigcup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu(A_i)$). Product measure

- Probability measure $\mathbb{P}$ is a measure such that $\mathbb{P}(\Omega) = 1$. Inclusion-exclusion principle ($\mathbb{P}(A_1 \cup A_2) = \mathbb{P}(A_1) + \mathbb{P}(A_2) - \mathbb{P}(A_1 \cap A_2)$ and general case $n \geq 2$).

- Conditional probability. $B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A|B) := \mathbb{P}(A \cap B)/\mathbb{P}(B)$. $\mathbb{P}(.|B)$ is a probability measure. Total probability rule (if $(A_i)_{i \in \mathbb{N}}$ is a partition of Ω, $\mathbb{P}(C) = \sum_{i=1}^{\infty} \mathbb{P}(C \cap A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)\mathbb{P}(C|A_i)$). Bayes rule: $\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}$.

- Independence of a collection of events $\{A_i,\ i \in S\}$ $\mathbb{P}(\bigcap_{i \in I} A_i) = \prod_{i \in I} \mathbb{P}(A_i)$ for any finite index set $I \subset S$. Independence does not imply conditional independence.

- Borel-Cantelli lemmas.

1. $\sum_{i=1}^{\infty} \mathbb{P}(A_i) < \infty \Rightarrow \mathbb{P}(A_i \ i.o) = 0$.
2. If $\{A_i, \ i = 1, 2, \ldots\}$ independent events and $\sum_{i=1}^{\infty} \mathbb{P}(A_i) = \infty$, then $\mathbb{P}(A_i \ i.o) = 1$.

2 Counting

Connection between counting and discrete uniform probability measure. Number of subsets, permutations, ordered and unordered arrangements.

Composition of $n \in \mathbb{N}$: how many ways to write n as a k-sum of non-negative integers (Recitation 3).

3 Integration

- Measurable functions from $(\Omega, \mathcal{F})$ to $(\Omega', \mathcal{F}')$. Construction of measurable functions by sum, product, composition, limit. Special case of random variables, i.e, $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B})$. $\{\omega | X(\omega) < c\} \in \mathcal{F}$. Continuity implies measurability.

- Abstract integration $\int g d\mu$, $\int_\Omega g(\omega) d\mu(\omega)$. Construction from simple functions to integrable functions (g measurable and $\int |g| d\mu < \infty$). For non-negative measurable functions, integral can be defined by $+\infty$. A series can be interpreted as an integral with respect to an impulse train (PS 5). Properties of integral: positivity ($\int f \geq 0$ for $f \geq 0$), monotonicity (if $f \leq g$ a.s., then $\int f d\mu \leq \int g d\mu$), linearity ($\int (\alpha f(\omega) + \beta g(\omega) d\mu(\omega)) = \alpha \int f d\mu + \beta \int g d\mu$), $\int |g| d\mu \geq |\int g d\mu|$.

- Theorems to interchange integration and limit: Fatou lemma ($\int \liminf f_n d\mu \leq \liminf \int f_n d\mu$), monotone convergence theorem (if $f_n \downarrow f$ a.s, then $\int f_n \downarrow \int f$), dominated convergence theorem (if $f_n \to f$, $|f_n| \leq g$ a.s. and $\int g < \infty$, then $\int f_n \to \int f$).

- Fubini's theorem to interchange integration order, and also integration and differentiation: $(\Omega_k, \mathcal{F}_k, \mu_k)$, $k = 1, 2$ be σ-finite measure space and $f(\omega_1, \omega_2)$ integrable on the product measure space with measure π. Assume any of the two conditions holds

  1. f measurable and $f \geq 0$ (in which case the integrals can be $+\infty$)
  2. f integrable

  Then $\int_{\omega_k} f(\omega_k, .) d\mu_k(\omega_k)$, $k = 1, 2$ are integrable functions of the other variable and $\int f d\pi = \int_{\omega_1} \int_{\omega_2} f(\omega_1, \omega_2) d\mu_2(\omega_2) d\mu_1(\omega_1) = \int_{\omega_2} \int_{\omega_1} f(\omega_1, \omega_2) d\mu_1(\omega_1) d\mu_2(\omega_2)$.

4 Random variables

- Let $\mathbb{P}$ probability measure on $(\Omega, \mathcal{F})$. A random variable X is a measurable function from $(\Omega, \mathcal{F})$ into $(\mathbb{R}, \mathcal{B})$, define the distribution function $F_X(x) = \mathbb{P}(\{\omega | X(\omega) \leq x\})$. F_X is increasing, right-continuous and range from zero to one. F_X characterizes completely the probability law of X: the induced probability measure $\mathbb{P}_X$ on $(\mathbb{R}, \mathcal{B})$ by $\mathbb{P}_X((-\infty, c]) = F_X(x)$. F_X can be used to simulate X from a uniform random variable on $[0,1]$. Oftentimes, it is easier to characterize the distribution function F_X of a random variable X than its probability law $\mathbb{P}_X$ (e.g. the derived distribution $Y = g(X)$ in PS 4).

  Discrete random variable indexed by countable set I: $X(\Omega) = \{x_i, \ i \in I\}$. $p_i := \mathbb{P}(X = x_i) \geq 0$ and $\sum_{i \in I} p_i = 1$. X is a continuous random variable with respect to the Lebesgue measure if there exists a measurable density $f_X \geq 0$ such that $F_X(x) = \int_{-\infty}^{x} f_X(t)dt$. f_X can be computed from F_X by $f_X = F'_X$.

- If X integrable, note $E[X] = \int X(\omega) d\mathbb{P}(\omega) = \int x dF_X(x)$. Law of the unconscious statistician for integrable g ($E[g(X)] = \int g(x) dF_X(x)$, discrete case: $E[g(X)] = \sum_{i \in I} p_i g(x_i)$, continuous case: $E[g(X)] = \int g(x) f_X(x) dx$. Jensen's inequality (if f convex, then $f(E[X]) \leq E[f(X)]$). If $X \geq 0$ and $E[X] < \infty$ then $X < \infty$ w.p. 1, if $X \geq 0$ and $E[X] = 0$ then $X = 0$ w.p 1. If X^2 is integrable, define the variance $var(X) = E[X^2] - E[X]^2$. $var(\alpha X + \beta) = \alpha^2 var(X)$. $\mathbb{P}(B) = E[\mathbf{I}_B]$, $\mathbf{I}_B$ is the indicator function of the set $B \in \mathcal{F}$.

- Random variables $X, Y, X_1, \ldots, X_n$ on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. $F_{X,Y}(x,y) = \mathbb{P}(X \leq x; Y \leq y)$.

  X, Y are jointly continuous if there exists a measurable density $f_{X,Y} \geq 0$ such that $F_{X,Y}(x,y) = \int_{s=-\infty}^{x} \int_{t=-\infty}^{y} f_{X,Y}(s,t) ds\ dt$. $f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}}{\partial x \partial y}(x,y)$ where $f_{X,Y}$ continuous.

  Jacobian rule for derived distributions by change of variables: if $X_1, \ldots, X_n$ are jointly continuous random variables with density $f_{X_1,\ldots,X_n}$ and $g : \mathbb{R}^n \to \mathbb{R}^n$ is a continuously differentiable invertible mapping, then $(Y_1, \ldots, Y_n) = g(X_1, \ldots, X_n)$ are jointly continuous random variables on the range of g with density $f_{Y_1,\ldots,Y_n}(y_1, \ldots, y_n) = f_{X_1,\ldots,X_n}(g^{-1}(y_1, \ldots, y_n))|J(y_1, \ldots, y_n)|$, where J is the Jacobian of g^{-1}, i.e., the determinant of the matrix $\frac{\partial g^{-1}}{\partial y}$.

- Bayesian inference $\mathbb{P}$. X, Y discrete, $\mathbb{P}_X(X = x|y) = \mathbb{P}(Y = y|x)\mathbb{P}(X = x)/\mathbb{P}(Y = y)$. If X, Y are jointly continuous and $f_Y(y) > 0$, we can define $f_{X|Y}(x|y) = f_{X,Y}(x,y)/f_Y(y)$. $f_{X|Y}(x|y) = f_{Y|X}(y|x) f_{X(x)}/f_Y(y)$. If X continuous and Y discrete, $f_{X|Y}(x|y) = \mathbb{P}(Y = y|x) f_{X(x)}/\mathbb{P}(Y = y)$.

  Conjugate distribution for inference (Recitation 9).

- Abstract definition of conditional expectation: there exists a measurable function ψ such that $E[Yg(X)] = E[g(X)\psi(X)]$ for all bounded measurable g. $E[Y|X] := \psi(X)$ is unique almost surely and is called conditional expectation. For jointly continuous random variables X, Y, $E[Y|X] = \int y f_{Y|X}(y|X)dy$.

Law of total expectation $E[Y] = E_X[E_Y[Y|X]]$. Law of total variance $var(Y) = E_X[var(Y|X)] + var_X(E[Y|X])$.

Conditional expectation has all the properties of the usual expectation. Conditional expectation with respect to a sub σ-field $\mathcal{H} \subset \mathcal{F}$: $E[X|\mathcal{H}] := E[X|\mathbf{I}_H,\ H \in \mathcal{H}]$. If Z is $\mathcal{H}$-measurable, $E[ZX|\mathcal{H}] = ZE[X|\mathcal{H}]$ a.s.

Tower property: $\mathcal{G} \subset \mathcal{H} \subset \mathcal{F}$ $E[E[X|\mathcal{H}]|\mathcal{G}] = E[X|\mathcal{G}]$.

- X, Y are independent, noted $X \perp Y$, if for all $x, y \in \mathbb{R}$, the two events $\{X \leq x\}$ and $\{Y \leq y\}$ are independent, i.e. $F_{X,Y}(x,y) = F_X(x)F_Y(y)$. If X, Y are jointly continuous, $f_{X,Y}(x,y) = f_X(x)f_Y(y)$ or equivalently $f_{X|Y}(x|y) = f_X(x)$ for all x, y such that $f_Y(y) > 0$. If X, Y are independent and continuous, they are jointly continuous with density $f_X(x)f_Y(y)$.

  If $X \perp Y$, $f(X) \perp g(Y)$ for any measurable f, g. If X, Y are continuous random variables and $X \perp Y$, $Z = X + Y$ is continuous with density given by the convolution $f_Z(z) = \int f_X(s)f_Y(z-s)ds$.

- Assume random variables are square-integrable. Covariance $cov(X,Y) = E[(X-E[X])(Y-E[Y])]$. $var(X_1 + \ldots + X_n) = \sum_{i=1}^n var(X_i) + 2\sum_{i<j} cov(X_i, X_j)$. Schwartz inequality $(E[XY]^2 \leq E[X^2]E[Y^2])$. Corollary: $|cov(X,Y)| \leq \sqrt{var(X)}\sqrt{var(Y)}$. Assuming Y not a.s zero, there is equality in the last two inequalities with iff there is $a \neq 0, b$ such that $X = aY + b$. If $E[XY] = E[X]E[Y]$, then $cov(X,Y) = 0$ and X, Y are said uncorrelated. If X, Y independent, then X, Y uncorrelated.

- Miscellaneous.

  Order statistics of iid random variables (Recitation 6)

  Probabilistic method (PS 2): to show an object exists, define a probabilistic model and show that with positive probability it generates the desired object.

  Use of indicator functions to solve problems (e.g. exercise 5 in PS 5).

5 Transforms of random variables

- Let X be a random variable. Moment generating function $M_X(s) = E[e^{sX}] \in \mathbb{R} \cup \{+\infty\}$. Characteristics function $\phi_X(t) = E[e^{itX}]$ defined everywhere and continuous at zero. If $Y = aX + b$, then $M_Z(s) = e^{sb}M_X(sa)$. If $X \perp Y$, $M_{X+Y} = M_X M_Y$.

- Inversion theorem: $M_X(s) = M_Y(s)$ for all $x \in (-\epsilon, \epsilon)$, then $F_X = F_Y$. Transforms give another representation of a distribution function. Know transform of common random variables, including discrete distribution, to toggle between representation.

- Continuity theorem: let ϕ_n be the characteristic function of some distribution function F_n and assume $\phi_n(t) \to \phi(t)$ for all t such that ϕ continuous at zero, then there exists a distribution function F such that with characteristics function ϕ and $F_n(t) \to F(t)$ for all t.

- Random sum of iid random variables X_i: $Y = \sum_{i=1}^{N} X_i$. $E[Y] = E[X]E[N]$. $M_Y(s)$ is the moment generating function of N with e^s replaced by $M_X(s)$.

- If M_X is finite on a neighborhood of 0, then X has moments at all orders and $E[X^k] = \frac{d^k M_X}{ds^k}(0)$.

6 Multivariate normal random variables

- A normal random variable $\mathcal{N}(m, V)$ with mean m and (positive semi-definite) covariance matrix $V \succeq 0$ has moment generating function $M(s) = \exp(s^T m)\exp(s^T V s/2)$. Let $(X_1, \ldots, X_n)$ be normal random vector. If the X_i's are uncorrelated, then they are independent.

- Density of a non-degenerate n-dimensional normal random variable with mean m and covariance matrix $V \succ 0$: $f(x) = \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2}(x-m)^T V^{-1}(x-m))$.

- If X is a normal random vector on $\mathbb{R}^n$ with mean m and covariance V and A a $m \times n$ matrix, $AX + b$ is normal random vector with $Am + b$ and covariance AVA^T.

- If (X, Y) is a zero-mean normal vector (and not only X, Y normal vectors) with covariance $\begin{pmatrix} V_{XX} & V_{XY} \\ V_{XY}^T & V_{YY} \end{pmatrix}$ and $V_{YY} \succ 0$, then $E[X|Y] = V_{XY} V_{YY}^{-1} Y$. Moreover, $X - E[X|Y]$ is a normal random vector independent of Y.

7 Concentration and limit theorems

- Markov inequality: if $h \geq 0$ integrable and $a > 0$, then $\mathbb{P}(h(X) \geq a) \leq \frac{E[h(X)]}{a}$. Chebyshev inequality $\mathbb{P}(|X - E[X]|^2 \geq \epsilon^2) \leq \frac{var(X)}{\epsilon^2}$.

- Different modes of convergence of a sequence of random variables (X_n) to a limit X. Convergence almost surely ($\mathbb{P}(|X_n(\omega) - X(\omega)| \to 0) = 1$), in probability (for all $\epsilon > 0$, $\mathbb{P}(|X_n - X| \geq \epsilon) \to 0$), in rth mean - $r \geq 1$- ($E[|X_n - X|^r] \to 0$), in distribution ($F_{X_n}(t) \to F_X(t)$ for all t where F_X continuous, or equivalently $E[g(X_n)] \to E[g(X)]$ for all bounded continuous g). Convergence in distribution does not involve the sample space Ω, nor any link between $X_n(\omega)$ and $X(\omega)$. However, if $X_n \xrightarrow{D} X$, then there exist a version Y_n of X_n on a sample space such that $Y_n \xrightarrow{a.s} Y$.

- Some convergence modes are stronger. The picture is $\xrightarrow{a.s} \quad \xrightarrow{i.p} \quad \xrightarrow{D}$ and for $q \geq r \geq 1$, $\xrightarrow{q} \Rightarrow \xrightarrow{r} \Rightarrow \xrightarrow{i.p}$.

  If there exists M such that $|X_n| \leq M$ a.s. and $X_n \xrightarrow{i.p} X$, then $X_n \xrightarrow{r} X$ for all $r \geq 1$. If X_n converge in distribution to a constant, then it converges in probability (PS 7).

- Be careful with incorrect "generalizations" of analysis such as X_n, Y_n have respective limits X, Y implies $X_n + Y_n$ has limit $X + Y$. Not all such statements hold in probability.

- Let $X_i,\ i = 1, 2, \ldots$ be iid random variables with $E[|X_1|] < \infty$ and $m = E[X_1]$.

  Weak law of large numbers: $\frac{1}{n}\sum_{i=1}^{n} X_i - m \rightarrow$ in probability.

  Strong law of large numbers: $\frac{1}{n}\sum_{i=1}^{n} X_i - m \rightarrow 0$ almost surely.

  In addition, assume $var(X_1) = \sigma^2 < \infty$, then Central limit theorem states $\sqrt{n}\frac{1}{n}\left(\sum_{i=1}^{n} X_i - m\right) \rightarrow \mathcal{N}(0, \sigma^2)$

8 Renewal processes

- Renewal processes is continuous-time arrival process started at zero with iid inter-arrival time. Discrete-time Bernouilli process.

- Poisson process with rate λ is a renewal process with inter-arrival time exponentially distributed with parameter λ. Memoryless property. Independence of disjoint time intervals. Given the second arrival time $Y_2 = t$, the first arrival time Y_1 has density s/t for $0 \leq s \leq t$.

- Merging of two independent Poisson processes with rate λ_1, λ_2 yields a Poisson process with rate $\lambda_1 + \lambda_2$. Splitting of a Poisson process with rate λ with probability p yields two independent Poisson processes with rate $p\lambda$ and $(1-p)\lambda$. Not true for discrete time Bernouilli process. When dealing with exponential or Erlang random variables, sometimes easier to build corresponding Poisson processes to analyze (Recitation 10).

- "Steady-state:" wait long enough to reach time-invariance of renewal process (which is started at time 0). Random incidence in renewal processes with inter-arrival time T: select an interval containing a "random" time point. Length X of the randomly selected interval has density $f_X(x) = \frac{x f_T(x)}{E[T]}$, $E[X] = E[T^2]/E[T]$.

9 Finite-state Markov chain

- Finite state space $\mathcal{S}$. A Markov chain is a sequence $X_0, X_1, \ldots$ of random variables on $\mathcal{S}$ with the Markov property $\mathbb{P}(X_{n+1} = j | X_n = i, \ldots, X_0) =$

$\mathbb{P}(X_{n+1} = j|X_n = i) = P^n_{ij}$. If homogeneous, no dependence on time n. N stopping rule: $\mathbf{I}_{N=n} = g_n(X_0, \ldots, X_n)$. For finite state, Markov chain has strong Markov property: for all stopping rule N, $\mathbb{P}(X_{N+1} = j|X_N, \ldots, X_0) = \mathbb{P}(X_{N+1} = j|X_N)$.

- $(P_{ij})_{(i,j)\in\mathcal{S}^2}$ stochastic matrix. n-step transition probability $r_{ij}(n) = \mathbb{P}(X_n = j|X_0 = i) = (P^n)_{ij}$. State i communicate with j if j visited with positive probability from i after some time. State i is recurrent if for every j such that $i \to j$, $j \to i$, in which case we say that i, j are in the same recurrent class. Starting in a recurrent class, all states of the class are visited infinitely often with probability one. There can be several recurrent classes. i transient state if there is a state j such that $i \to j$ but $j \nrightarrow i$. Then $r_{ki} \to 0$ for all $k \in \mathcal{S}$. The system eventually leaves all transient states and is absorbed to a recurrent class geometrically quickly.

- First passage time to a state and mean recurrence time t^*_i within a recurrent class. Bellman equations. Beware of modified Markov chain given that a state is not visited (PS 9, Problem 2 and 3).

- Invariant distribution $\pi \geq 0$: $\pi^T\mathbf{1} = 1$ and $\pi^T = \pi^T P$ (balance equation). Always exists, but not always unique. Gives zero mass to transient states. $\pi_i = 1/t^*_i$ for i in a recurrent class and 0 elsewhere is an invariant distribution. Unique if only one recurrent class. Birth-death process. Balance equations from cut in the Markov chain graph (Notes on Recitation 11).

- Period d_i of state i is $gcd(n|r_{ii}(n) > 0)$. If i has a self-transition, then i is aperiodic. Within a recurrent class $\mathcal{R}$, all states have same period. If $\mathcal{R}$ is aperiodic, there exists n_0 such that $r_{ij}(n) > 0$ for all $n \geq n_0$ and $i, j \in \mathcal{R}$. If one single recurrent class and aperiodic, then $\lim_n r_{ij}(n) = \pi_j$ for all $i, j \in \mathcal{S}$.

- Strong law of large number of aperiodic single recurrent class Markov chain (renewal reward theory). Let $N_t(i)$be the number of visits to state i in first t stages. Ensemble average $E\left[\frac{N_t(i)}{t}\right] \to \pi_i$. Time average: $\frac{N_t(i)}{t} \to \pi_i$ a.s.