Basic Probability Theory
Outline

- Definitions and theorems: independence, Bayes, ...
- Random variables: pdf, expectation, variance, typical distributions, ...
- Bounds: Markov, Chebyshev and Chernoff
- Method of indicators
- Multi-dimensional random variables: joint distribution, covariance, ...
- Maximum likelihood estimation
- Convergence: Central limit theorem and interesting limits
Elements of Probability

Definition:
- Sample Space $\Omega$: Set of all possible outcomes
- Event Space $\mathcal{F}$: A family of subsets of $\Omega$
- Probability Measure: Function $P : \mathcal{F} \rightarrow \mathbb{R}$ with properties:
  1. $P(A) \geq 0 \ (\forall A \in \mathcal{F})$
  2. $P(\Omega) = 1$
  3. $A_i$'s disjoint, then $P(\bigcup_i A_i) = \sum_i P(A_i)$

Sample spaces can be discrete (rolling a die) or continuous (wait time in line)
Conditional Probability and Independence

Conditional probability:

- For events $A$, $B$:

\[
P(A|B) = \frac{P(A \cap B)}{P(B)}
\]

- Intuitively means “probability of $A$ when $B$ is known”

Independence

- $A$, $B$ independent if $P(A|B) = P(A)$ or equivalently:

\[
P(A \cap B) = P(A)P(B)
\]

- Beware of intuition: roll two dice ($x_a$ and $x_b$), outcomes

\[
\{x_a = 2\} \text{ and } \{x_a + x_b = k\}
\]

are independent if $k = 7$, but not otherwise!
Basic laws and bounds

- Union bound: since $P(A \cup B) = P(A) + P(B) - P(A \cap B)$, we have

  $$P(\bigcup_i A_i) \leq \sum_i P(A_i)$$

- Law of total probability: if $\bigcup_i A_i = \Omega$, then

  $$P(B) = \sum_i P(A_i \cap B) = \sum_i P(A_i)P(B|A_i)$$

- Chain rule: $P(A_1, A_2, \ldots, A_N) = P(A_1)P(A_2|A_1)P(A_3|A_1, A_2) \cdots P(A_N|A_1, \ldots, A_{N-1})$

- Bayes rule: $P(A|B) = P(B|A) \frac{P(A)}{P(B)}$ (several versions)
Random Variables and Distributions

- A random variable $X$ is a function $X : \Omega \rightarrow \mathbb{R}$

  Example: Number of heads in 20 tosses of a coin

- Probabilities of events associated with random variables defined based on the original probability function. e.g.,
  
  $P(X = k) = P(\{\omega \in \Omega | X(\omega) = k\})$

- Cumulative Distribution Function (CDF) $F_X : \mathbb{R} \rightarrow [0, 1]$:
  
  $F_X(x) = P(X \leq x)$

- (X discrete) Probability Mass Function (pmf):
  
  $p_X(x) = P(X = x)$

- (X continuous) Probability Density Function (pdf):
  
  $f_X(x) = dF_X(x)/dx$
Properties of Distribution Functions

- **CDF:**
  - $0 \leq F_X(x) \leq 1$
  - $F_X$ monotone increasing, with $\lim_{x \to -\infty} F_X(x) = 0$, $\lim_{x \to \infty} F_X(x) = 1$

- **pmf:**
  - $0 \leq p_X(x) \leq 1$
  - $\sum_x p_X(x) = 1$
  - $\sum_{x \in A} p_X(x) = p_X(A)$

- **pdf:**
  - $f_X(x) \geq 0$
  - $\int_{-\infty}^{\infty} f_X(x) \, dx = 1$
  - $\int_{x \in A} f_X(x) \, dx = P(X \in A)$
Expectation and Variance

- Assume random variable $X$ has pdf $f_X(x)$, and $g : \mathbb{R} \rightarrow \mathbb{R}$. Then

$$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)\,dx$$

- for discrete $X$, $E[g(X)] = \sum_x g(x)p_X(x)$

- Expectation is linear:
  - for any constant $a \in \mathbb{R}$, $E[a] = a$
  - $E[ag(X)] = aE[g(X)]$
  - $E[g(X) + h(X)] = E[g(X)] + E[h(X)]$

Conditional Expectation

- $E[g(X, Y)| Y = a] = \sum_x g(x, a)p_{X|Y=a}(x)$ (similar for continuous random variables)

- Iterated expectation:

  $$E[g(X, Y)] = E_a[E[g(X, Y)| Y = a]]$$

Often useful in practice. Example: number of heads in $N$ flips of a coin with random bias $p \in [0, 1]$ with pdf $f_p(x) = 2(1 - x)$ is $\frac{N}{3}$
Some Common Random Variables

- \( X \sim \text{Bernoulli}(p) \) \((0 \leq p \leq 1): p_X(x) = \begin{cases} p & x = 1, \\ 1 - p & x = 0. \end{cases} \)

- \( X \sim \text{Geometric}(p) \) \((0 \leq p \leq 1): p_X(x) = p(1 - p)^{x-1} \)

- \( X \sim \text{Uniform}(a, b) \) \((a < b): f_X(x) = \begin{cases} \frac{1}{b - a} & a \leq x \leq b, \\ 0 & \text{otherwise}. \end{cases} \)

- \( X \sim \text{Normal}(\mu, \sigma^2): f_X(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{1}{2\sigma^2}(x-\mu)^2} \)
Binomial distribution

- Combinatorics: consider a bag with $n$ different balls
  - number of different ordered subsets with $k$ elements:
    \[ n(n - 1) \cdots (n - k + 1) \]
  - number of different unordered subsets with $k$ elements:
    \[ \binom{n}{k} = \frac{n!}{k!(n - k)!} \]
- $X \sim \text{Binomial}(n, p)$ ($n > 0$, $0 \leq p \leq 1$):
  \[ p_X(x) = \binom{n}{x} p^x (1 - p)^{n-x} \]
Method of indicators

- Goal: find expected number of successes out of $N$ trials
- Method: define an indicator (Bernoulli) random variable for each trial, find expected value of the sum
- Examples:
  - Bowl with $N$ spaghetti strands. Keep picking ends and joining. Expected number of loops?
  - $N$ drunk sailors pass out on random bunks. Expected number on their own?
Some Useful Inequalities

- Markov’s Inequality: If $X$ is a random variable, and $a > 0$. Then:
  \[ P(|X| \geq a) \leq \frac{E[|X|]}{a} \]

- Chebyshev’s Inequality: If $E[X] = \mu$, $Var(X) = \sigma^2$, $k > 0$, then:
  \[ Pr(|X - \mu| \geq k\sigma) \leq \frac{1}{k^2} \]

- Chernoff bound: Let $X_1, \ldots, X_n$ independent Bernoulli with $P(X_i = 1) = p_i$. Denoting $\mu = E[\sum_{i=1}^{n} X_i] = \sum_{i=1}^{n} p_i$,
  \[ P(\sum_{i=1}^{n} X_i \geq (1 + \delta)\mu) \leq \left( \frac{e^\delta}{(1 + \delta)^{1+\delta}} \right)^\mu \]

  for any $\delta$. Multiple variants of Chernoff-type bounds exist, which can be useful in different settings.
Multiple Random Variables and Joint Distributions

$X_1, \ldots, X_n$ random variables

- Joint CDF: $F_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = P(X_1 \leq x_1, \ldots, X_n \leq x_n)$
- Joint pdf: $f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) = \frac{\partial^n F_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{\partial x_1 \ldots \partial x_n}$
- Marginalization: 
  
  $f_{X_1}(x_1) = \int_{-\infty}^{\infty} \ldots \int_{-\infty}^{\infty} f_{X_1,\ldots,X_n}(x_1, \ldots, x_n) dx_2 \ldots dx_n$

- Conditioning: $f_{X_1|X_2,\ldots,X_n}(x_1|x_2, \ldots, x_n) = \frac{f_{X_1,\ldots,X_n}(x_1,\ldots,x_n)}{f_{X_2,\ldots,X_n}(x_2,\ldots,x_n)}$

- Chain Rule: $f(x_1, \ldots, x_n) = f(x_1) \prod_{i=2}^{n} f(x_i|x_1, \ldots, x_{i-1})$

- Independence: $f(x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i)$. 
Random Vectors

\(X_1, \ldots, X_n\) random variables. \(X = [X_1 X_2 \ldots X_n]^T\) random vector.

- If \(g : \mathbb{R}^n \rightarrow \mathbb{R}\), then
  \[E[g(X)] = \int_{\mathbb{R}^n} g(x_1, \ldots, x_n)f_{X_1,\ldots,X_n}(x_1, \ldots, x_n)\,dx_1 \ldots dx_n\]
- If \(g : \mathbb{R}^n \rightarrow \mathbb{R}^m\), \(g = [g_1 \ldots g_m]^T\), then
  \[E[g(X)] = [E[g_1(X)] \ldots E[g_m(X)]]^T\]

- Covariance Matrix:
  \[\Sigma = \text{Cov}(X) = E[(X - E[X])(X - E[X])^T]\]

- Properties of Covariance Matrix:
  \(\Sigma_{ij} = \text{Cov}[X_i, X_j] = E[(X_i - E[X_i])(X_j - E[X_j])]\)
  \(\Sigma\) symmetric, positive semidefinite
Multivariate Gaussian Distribution

\( \mu \in \mathbb{R}^n, \Sigma \in \mathbb{R}^{n \times n} \) symmetric, positive semidefinite
\( X \sim \mathcal{N}(\mu, \Sigma) \) n-dimensional Gaussian distribution:

\[
f_X(x) = \frac{1}{(2\pi)^{n/2} \det(\Sigma)^{1/2}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right)
\]

- \( E[X] = \mu \)
- \( \text{Cov}(X) = \Sigma \)
Parameter Estimation: Maximum Likelihood

- Parametrized distribution $f_X(x; \theta)$ with parameter(s) $\theta$ unknown.
- IID samples $x_1, \ldots, x_n$ observed.
- Goal: Estimate $\theta$
- (Ideally) MAP: $\hat{\theta} = \arg\max_\theta \{ f_{\Theta|X}(\theta|X = (x_1, \ldots, x_n)) \}$
- (In practice) MLE: $\hat{\theta} = \arg\max_\theta \{ f_{X|\theta}(x_1, \ldots, x_n; \theta) \}$
**MLE Example**

\( X \sim \text{Gaussian}(\mu, \sigma^2) \). \( \theta = (\mu, \sigma^2) \) unknown. Samples \( x_1, \ldots, x_n \).

Then:

\[
f(x_1, \ldots, x_n; \mu, \sigma^2) = \left( \frac{1}{2\pi\sigma^2} \right)^{n/2} \exp \left( - \frac{\sum^n_{i=1} (x_i - \mu)^2}{2\sigma^2} \right)
\]

Setting: \( \frac{\partial \log f}{\partial \mu} = 0 \) and \( \frac{\partial \log f}{\partial \sigma} = 0 \)

Gives:

\[
\hat{\mu}_{\text{MLE}} = \frac{\sum^n_{i=1} x_i}{n}, \quad \hat{\sigma}^2_{\text{MLE}} = \frac{\sum^n_{i=1} (x_i - \hat{\mu})^2}{n}
\]

Sometimes it is not possible to find the optimal estimate in closed form, then iterative methods can be used.
Central limit theorem:

- Central limit theorem: Let $X_1, X_2, \ldots, X_n$ be iid with finite mean $\mu$ and finite variance $\sigma^2$, then the random variable $Y = \frac{1}{n} \sum_{i=1}^{n} X_i$ is approximately Gaussian with mean $\mu$ and variance $\frac{\sigma^2}{n}$.
- Approximation becomes better as $n$ grows.
- Law of large numbers as a corollary.
Interesting limits

- \( \lim_{n \to \infty} (1 + \frac{k}{n})^n \to e^k \)
- \( \lim_{n \to \infty} n! \to \sqrt{2\pi n} \left( \frac{n}{e} \right)^n \) (lower bound)
- \( \lim_{n \to \infty} n^{\frac{1}{n}} \to 1 \)
- \( \lim_{(n, \epsilon) \to (\infty, 0)} \text{Binomial}(n, \epsilon) \to \text{Poisson}(n\epsilon) \)
- \( \lim_{n \to \infty} \text{Binomial}(n, p) \to \text{Normal}(np, np(1 - p)) \)
References

1. CS229 notes on basic linear algebra and probability theory
2. Wikipedia!