Probability Review

Gonzalo Mateos
Dept. of ECE and Goergen Institute for Data Science
University of Rochester
gmateosb@ece.rochester.edu
http://www.ece.rochester.edu/~gmateosb/

September 1, 2016
Markov and Chebyshev’s inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation
Markov’s inequality

- RV $X$ with $E[|X|] < \infty$, constant $a > 0$
- Markov’s inequality states $\Rightarrow P(|X| \geq a) \leq \frac{E(|X|)}{a}$

Proof.

- $\mathbb{I}\{|X| \geq a\} = 1$ when $|X| \geq a$ and 0 else. Then (figure to the right)
  \[ a\mathbb{I}\{|X| \geq a\} \leq |X| \]

- Use linearity of expected value
  \[ aE(\mathbb{I}\{|X| \geq a\}) \leq E(|X|) \]

- Indicator function’s expectation = Probability of indicated event
  \[ aP(|X| \geq a) \leq E(|X|) \]
Chebyshev’s inequality

- RV $X$ with $\mathbb{E}(X) = \mu$ and $\mathbb{E}[(X - \mu)^2] = \sigma^2$, constant $k > 0$
- Chebyshev’s inequality states $\Rightarrow P(|X - \mu| \geq k) \leq \frac{\sigma^2}{k^2}$

Proof.
- Markov’s inequality for the RV $Z = (X - \mu)^2$ and constant $a = k^2$

$$P((X - \mu)^2 \geq k^2) = P(|Z| \geq k^2) \leq \frac{\mathbb{E}[|Z|]}{k^2} = \frac{\mathbb{E}[(X - \mu)^2]}{k^2}$$

- Notice that $(X - \mu)^2 \geq k^2$ if and only if $|X - \mu| \geq k$ thus

$$P(|X - \mu| \geq k) \leq \frac{\mathbb{E}[(X - \mu)^2]}{k^2}$$

- Chebyshev’s inequality follows from definition of variance
If absolute expected value is finite, i.e., $E[|X|] < \infty$

$\Rightarrow$ Complementary (c)cdf decreases at least like $x^{-1}$ (Markov’s)

If mean $E(X)$ and variance $E[(X - \mu)^2]$ are finite

$\Rightarrow$ Ccdf decreases at least like $x^{-2}$ (Chebyshev’s)

Most cdfs decrease exponentially (e.g. $e^{-x^2}$ for normal)

$\Rightarrow$ Power law bounds $\propto x^{-\alpha}$ are loose but still useful

Markov’s inequality often derived for nonnegative RV $X \geq 0$

$\Rightarrow$ Can drop the absolute value to obtain $P(X \geq a) \leq \frac{E(X)}{a}$

$\Rightarrow$ General bound $P(X \geq a) \leq \frac{E(X^r)}{a^r}$ holds for $r > 0$
Convergence of random variables

Markov and Chebyshev’s inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation
Limits

- Sequence of RVs $X_N = X_1, X_2, \ldots, X_n, \ldots$
  - Distinguish between random process $X_N$ and realizations $x_N$

Q1) Say something about $X_n$ for $n$ large? $\Rightarrow$ Not clear, $X_n$ is a RV

Q2) Say something about $x_n$ for $n$ large? $\Rightarrow$ Certainly, look at $\lim_{n \to \infty} x_n$

Q3) Say something about $P (X_n \in \mathcal{X})$ for $n$ large? $\Rightarrow$ Yes, $\lim_{n \to \infty} P (X_n \in \mathcal{X})$

- Translate what we now about regular limits to definitions for RVs

- Can start from convergence of sequences: $\lim_{n \to \infty} x_n$
  - Sure and almost sure convergence

- Or from convergence of probabilities: $\lim_{n \to \infty} P (X_n)$
  - Convergence in probability, in mean square and distribution
Convergence of sequences and sure convergence

- Denote sequence of numbers \( x_N = x_1, x_2, \ldots, x_n, \ldots \)

- **Def:** Sequence \( x_N \) converges to the value \( x \) if given any \( \epsilon > 0 \)
  \( \Rightarrow \) There exists \( n_0 \) such that for all \( n > n_0 \), \( |x_n - x| < \epsilon \)

- Sequence \( x_n \) comes arbitrarily close to its limit \( \Rightarrow |x_n - x| < \epsilon \)
  \( \Rightarrow \) And stays close to its limit for all \( n > n_0 \)

- Random process (sequence of RVs) \( X_N = X_1, X_2, \ldots, X_n, \ldots \)
  \( \Rightarrow \) Realizations of \( X_N \) are sequences \( x_N \)

- **Def:** We say \( X_N \) converges surely to RV \( X \) if
  \( \Rightarrow \lim_{n \to \infty} x_n = x \) for all realizations \( x_N \) of \( X_N \)

- Said differently, \( \lim_{n \to \infty} X_n(s) = X(s) \) for all \( s \in S \)

- **Not really adequate.** Even a (practically unimportant) outcome that happens with vanishingly small probability prevents sure convergence
Almost sure convergence

- RV $X$ and random process $X_N = X_1, X_2, \ldots, X_n, \ldots$
- **Def:** We say $X_N$ converges almost surely to RV $X$ if
  \[ P \left( \lim_{n \to \infty} X_n = X \right) = 1 \]
  \[ \Rightarrow \text{Almost all sequences converge, except for a set of measure 0} \]
- Almost sure convergence denoted as
  \[ \lim_{n \to \infty} X_n = X \text{ a.s.} \]
  \[ \Rightarrow \text{Limit } X \text{ is a random variable} \]

**Example**
- $X_0 \sim \mathcal{N}(0, 1)$ (normal, mean 0, variance 1)
- $Z_n$ sequence of Bernoulli RVs, parameter $p$
- Define \[ \Rightarrow X_n = X_0 - \frac{Z_n}{n} \]
- $\frac{Z_n}{n} \to 0$ so \[ \lim_{n \to \infty} X_n = X_0 \text{ a.s. (also surely)} \]
Convergence in probability

- **Def:** We say $X_N$ converges in probability to RV $X$ if for any $\epsilon > 0$

  $$\lim_{n \to \infty} P(|X_n - X| < \epsilon) = 1$$

  $\Rightarrow$ Prob. of distance $|X_n - X|$ becoming smaller than $\epsilon$ tends to 1

- Statement is about probabilities, not about realizations (sequences)
  $\Rightarrow$ Probability converges, realizations $x_N$ may or may not converge
  $\Rightarrow$ Limit and prob. interchanged with respect to a.s. convergence

**Theorem**

*Almost sure (a.s.) convergence implies convergence in probability*

**Proof.**

- If $\lim_{n \to \infty} X_n = X$ then for any $\epsilon > 0$ there is $n_0$ such that

  $$|X_n - X| < \epsilon \text{ for all } n \geq n_0$$

- True for all almost all sequences so $P(|X_n - X| < \epsilon) \to 1$
Convergence in probability example

- $X_0 \sim \mathcal{N}(0, 1)$ (normal, mean 0, variance 1)
- $Z_n$ sequence of Bernoulli RVs, parameter $1/n$
- Define $\Rightarrow X_n = X_0 - Z_n$
- $X_n$ converges in probability to $X_0$ because
  \[
P(|X_n - X_0| < \epsilon) = P(|Z_n| < \epsilon) = 1 - P(Z_n = 1) = 1 - \frac{1}{n} \to 1\]
- Plot of path $x_n$ up to $n = 10^2, \ n = 10^3, \ n = 10^4$
  $\Rightarrow Z_n = 1$ becomes ever rarer but still happens
Almost sure convergence implies that almost all sequences converge.

Convergence in probability does not imply convergence of sequences.

Latter example: $X_n = X_0 - Z_n$, $Z_n$ is Bernoulli with parameter $1/n$

⇒ Showed it converges in probability

$$P \left( |X_n - X_0| < \epsilon \right) = 1 - \frac{1}{n} \to 1$$

⇒ But for almost all sequences, $\lim_{n \to \infty} x_n$ does not exist.

Almost sure convergence ⇒ disturbances stop happening.

Convergence in prob. ⇒ disturbances happen with vanishing freq.

Difference not irrelevant

⇒ Interpret $Z_n$ as rate of change in savings

⇒ With a.s. convergence risk is eliminated

⇒ With convergence in prob. risk decreases but does not disappear
Mean-square convergence

- **Def:** We say $X_N$ converges in mean square to RV $X$ if

\[
\lim_{n \to \infty} \mathbb{E} \left[ |X_n - X|^2 \right] = 0
\]

$\Rightarrow$ Sometimes (very) easy to check

**Theorem**

*Convergence in mean square implies convergence in probability*

**Proof.**

- From Markov's inequality

\[
P \left( |X_n - X| \geq \epsilon \right) = P \left( |X_n - X|^2 \geq \epsilon^2 \right) \leq \frac{\mathbb{E} \left[ |X_n - X|^2 \right]}{\epsilon^2}
\]

- If $X_n \to X$ in mean-square sense, $\mathbb{E} \left[ |X_n - X|^2 \right]/\epsilon^2 \to 0$ for all $\epsilon$

- Almost sure and mean square $\Rightarrow$ neither one implies the other
Convergence in distribution

- Consider a random process \( X_N \). Cdf of \( X_n \) is \( F_n(x) \)
- **Def:** We say \( X_N \) converges in distribution to RV \( X \) with cdf \( F_X(x) \) if
  \[ \lim_{n \to \infty} F_n(x) = F_X(x) \]
  for all \( x \) at which \( F_X(x) \) is continuous
- No claim about individual sequences, just the cdf of \( X_n \)
  \[ \Rightarrow \text{Weakest form of convergence covered} \]
- Implied by almost sure, in probability, and mean square convergence

**Example**

- \( Y_n \sim \mathcal{N}(0,1) \)
- \( Z_n \) Bernoulli with parameter \( p \)
- Define \( \Rightarrow X_n = Y_n - 10Z_n/n \)
- \( \frac{Z_n}{n} \to 0 \) so
  \[ \lim_{n \to \infty} F_n(x) "=" \mathcal{N}(0,1) \]
Convergence in distribution (continued)

- Individual sequences $x_n$ do not converge in any sense
  \[ \Rightarrow \text{It is the distribution that converges} \]

- As the effect of $Z_n/n$ vanishes pdf of $X_n$ converges to pdf of $Y_n$
  \[ \Rightarrow \text{Standard normal } \mathcal{N}(0,1) \]
Implications

- Sure $\Rightarrow$ almost sure $\Rightarrow$ in probability $\Rightarrow$ in distribution
- Mean square $\Rightarrow$ in probability $\Rightarrow$ in distribution
- In probability $\Rightarrow$ in distribution
Limit theorems

Markov and Chebyshev’s inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation
Introduction to Random Processes Probability Review 18

Sum of independent identically distributed RVs

- Independent identically distributed (i.i.d.) RVs $X_1, X_2, \ldots, X_n, \ldots$
- Mean $\mathbb{E}[X_n] = \mu$ and variance $\mathbb{E}[(X_n - \mu)^2] = \sigma^2$ for all $n$
- Q: What happens with sum $S_N := \sum_{n=1}^{N} X_n$ as $N$ grows?

- Expected value of sum is $\mathbb{E}[S_N] = N\mu \Rightarrow$ Diverges if $\mu \neq 0$
- Variance is $\mathbb{E}[(S_N - N\mu)^2] = N\sigma^2$
  $\Rightarrow$ Diverges if $\sigma \neq 0$ (always true unless $X_n$ is a constant, boring)

- One interesting normalization $\Rightarrow \bar{X}_N := (1/N) \sum_{n=1}^{N} X_n$
- Now $\mathbb{E}[\bar{X}_N] = \mu$ and $\text{var}[\bar{X}_N] = \sigma^2/N$
  $\Rightarrow$ Law of large numbers (weak and strong)

- Another interesting normalization $\Rightarrow Z_N := \frac{\sum_{n=1}^{N} X_n - N\mu}{\sigma \sqrt{N}}$
- Now $\mathbb{E}[Z_N] = 0$ and $\text{var}[Z_N] = 1$ for all values of $N$
  $\Rightarrow$ Central limit theorem
Law of large numbers

- Sequence of i.i.d. RVs $X_1, X_2, \ldots, X_n, \ldots$ with mean $\mu$
- Define sample average $\bar{X}_N := \frac{1}{N} \sum_{n=1}^{N} X_n$

**Theorem (Weak law of large numbers)**

*Sample average $\bar{X}_N$ of i.i.d. sequence converges in prob. to $\mu = \mathbb{E}[X_n]$*

$$\lim_{N \to \infty} P \left( |\bar{X}_N - \mu| < \epsilon \right) = 1, \quad \text{for all } \epsilon > 0$$

**Theorem (Strong law of large numbers)**

*Sample average $\bar{X}_N$ of i.i.d. sequence converges a.s. to $\mu = \mathbb{E}[X_n]$*

$$P \left( \lim_{N \to \infty} \bar{X}_N = \mu \right) = 1$$

- **Strong law implies weak law.** Can forget weak law if so wished
Proof of weak law of large numbers

- **Weak** law of large numbers is very simple to prove

Proof.

- Variance of $\bar{X}_n$ vanishes for $N$ large

$$\text{var} \left[ \bar{X}_N \right] = \frac{1}{N^2} \sum_{n=1}^{N} \text{var} \left[ X_n \right] = \frac{\sigma^2}{N} \to 0$$

- But, what is the variance of $\bar{X}_N$?

$$0 \leftarrow \frac{\sigma^2}{N} = \text{var} \left[ \bar{X}_N \right] = \mathbb{E} \left[ (\bar{X}_n - \mu)^2 \right]$$

- Then, $\bar{X}_N$ converges to $\mu$ in mean-square sense
  - Which implies convergence in probability

- **Strong** law is a little more challenging. Will not prove it here
Theorem (Central limit theorem)

Consider a sequence of i.i.d. RVs $X_1, X_2, \ldots, X_n, \ldots$ with mean $\mathbb{E}[X_n] = \mu$ and variance $\mathbb{E}[(X_n - \mu)^2] = \sigma^2$ for all $n$. Then

$$\lim_{N \to \infty} P \left( \frac{\sum_{n=1}^{N} X_n - N \mu}{\sigma \sqrt{N}} \leq x \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-u^2/2} \, du$$

- Former statement implies that for $N$ sufficiently large

$$Z_N := \frac{\sum_{n=1}^{N} X_n - N \mu}{\sigma \sqrt{N}} \sim \mathcal{N}(0, 1)$$

$\Rightarrow$ $Z_N$ converges in distribution to a standard normal RV

$\Rightarrow$ Remarkable universality. Distribution of $X_n$ arbitrary
Equivalently can say \[ \sum_{n=1}^{N} X_n \sim N(N\mu, N\sigma^2) \]

Sum of large number of i.i.d. RVs has a normal distribution
\[ \Rightarrow \text{Cannot take a meaningful limit here} \]
\[ \Rightarrow \text{But intuitively, this is what the CLT states} \]

Example

- Binomial RV \( X \) with parameters \((n, p)\)
- Write as \( X = \sum_{i=1}^{n} X_i \) with \( X_i \) i.i.d. Bernoulli with parameter \( p \)
- Mean \( \mathbb{E}[X_i] = p \) and variance \( \text{var}[X_i] = p(1 - p) \)
  \[ \Rightarrow \text{For sufficiently large } n \Rightarrow X \sim N(np, np(1 - p)) \]
Conditional probabilities

Markov and Chebyshev’s inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation
Recall definition of conditional probability for events $E$ and $F$

$$P(E \mid F) = \frac{P(E \cap F)}{P(F)}$$

⇒ Change in likelihoods when information is given, renormalization

**Def:** Conditional pmf of RV $X$ given $Y$ is (both RVs discrete)

$$p_{X \mid Y}(x \mid y) = P(X = x \mid Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}$$

Which we can rewrite as

$$p_{X \mid Y}(x \mid y) = \frac{P(X = x, Y = y)}{P(Y = y)} = \frac{p_{XY}(x, y)}{p_Y(y)}$$

⇒ Pmf for RV $X$, given parameter $y$ (“$Y$ not random anymore”)

**Def:** Conditional cdf is (a range of $X$ conditioned on a value of $Y$)

$$F_{X \mid Y}(x \mid y) = P(X \leq x \mid Y = y) = \sum_{z \leq x} p_{X \mid Y}(z \mid y)$$
Consider independent Bernoulli RVs $Y$ and $Z$, define $X = Y + Z$

**Q:** Conditional pmf of $X$ given $Y$? For $X = 0$, $Y = 0$

$$p_{X|Y}(X = 0 \mid Y = 0) = \frac{P(X = 0, Y = 0)}{P(Y = 0)} = \frac{(1 - p)^2}{1 - p} = 1 - p$$

**Or, from joint and marginal pmfs (just a matter of definition)**

$$p_{X|Y}(X = 0 \mid Y = 0) = \frac{p_{XY}(0, 0)}{p_Y(0)} = \frac{(1 - p)^2}{1 - p} = 1 - p$$

**Can compute the rest analogously**

$$p_{X|Y}(0 \mid 0) = 1 - p, \quad p_{X|Y}(1 \mid 0) = p, \quad p_{X|Y}(2 \mid 0) = 0$$
$$p_{X|Y}(0 \mid 1) = 0, \quad p_{X|Y}(1 \mid 1) = 1 - p, \quad p_{X|Y}(2 \mid 1) = p$$
Conditioning on sum of Poisson RVs

- Consider independent Poisson RVs $Y$ and $Z$, parameters $\lambda_1$ and $\lambda_2$
- Define $X = Y + Z$. Q: Conditional pmf of $Y$ given $X$?

$$p_{Y|X}(Y = y \mid X = x) = \frac{P(Y = y, X = x)}{P(X = x)} = \frac{P(Y = y)P(Z = x - y)}{P(X = x)}$$

- Used $Y$ and $Z$ independent. Now recall $X$ is Poisson, $\lambda = \lambda_1 + \lambda_2$

$$p_{Y|X}(Y = y \mid X = x) = \frac{e^{-\lambda_1} \lambda_1^y}{y!} \frac{e^{-\lambda_2} \lambda_2^{x-y}}{(x-y)!} \left[ \frac{e^{-(\lambda_1+\lambda_2)}(\lambda_1 + \lambda_2)^x}{x!} \right]^{-1}$$

$$= \frac{x!}{y!(x-y)!}(\frac{\lambda_1}{\lambda_1 + \lambda_2})^y (\frac{\lambda_2}{\lambda_1 + \lambda_2})^{x-y}$$

$$= \binom{x}{y} \left( \frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^y \left( \frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{x-y}$$

$\Rightarrow$ Conditioned on $X = x$, $Y$ is binomial $(x, \lambda_1/(\lambda_1 + \lambda_2))$
Def: Conditional pdf of RV $X$ given $Y$ is (both RVs continuous)

$$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)}$$

For motivation, define intervals $\Delta x = [x, x+dx]$ and $\Delta y = [y, y+dy]

⇒ Approximate conditional probability $P \left( X \in \Delta x \mid Y \in \Delta y \right)$ as

$$P \left( X \in \Delta x \mid Y \in \Delta y \right) = \frac{P(X \in \Delta x, Y \in \Delta y)}{P(Y \in \Delta y)} \approx \frac{f_{XY}(x, y)dx dy}{f_Y(y)dy}$$

⇒ From definition of conditional pdf it follows

$$P \left( X \in \Delta x \mid Y \in \Delta y \right) \approx f_{X|Y}(x \mid y)dx$$

⇒ What we would expect of a density

Def: Conditional cdf is

$$F_{X|Y}(x) = \int_{-\infty}^{x} f_{X|Y}(u \mid y)du$$
Communications channel example

- Random message (RV) $Y$, transmit signal $y$ (realization of $Y$)
- Received signal is $x = y + z$ ($z$ realization of random noise)
  - Model communication system as a relation between RVs
    $$X = Y + Z$$
  - Model additive noise as $Z \sim \mathcal{N}(0, \sigma^2)$ independent of $Y$
- Q: Conditional pdf of $X$ given $Y$? Try the definition
  $$f_{X|Y}(x \mid y) = \frac{f_{XY}(x, y)}{f_Y(y)} = \ ?$$
  - Problem is we don’t know $f_{XY}(x, y)$. Have to calculate
- Computing conditional probs. typically easier than computing joints
Communications channel example (continued)

- If $Y = y$ is given, then “$Y$ not random anymore”
  - It is still random in reality, we are thinking of it as given

- If $Y$ were not random, say $Y = y$ with $y$ given then $X = y + Z$
  - Cdf of $X$ given $Y = y$ now easy (use $Y$ and $Z$ independent)
    $\Pr(X \leq x \mid Y = y) = \Pr(y + Z \leq x \mid Y = y) = \Pr(Z \leq x - y)$

- But since $Z$ is normal with zero mean and variance $\sigma^2$
  $$\Pr(X \leq x \mid Y = y) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x-y} e^{-z^2/2\sigma^2} \, dz$$
  $$= \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} e^{-(z-y)^2/2\sigma^2} \, dz$$
  - $X$ given $Y = y$ is normal with mean $y$ and variance $\sigma^2$
Digital communications channel

- Conditioning is a common tool to compute probabilities

- Message 1 (w.p. \( p \)) \( \Rightarrow \) Transmit \( Y = 1 \)
- Message 2 (w.p. \( q \)) \( \Rightarrow \) Transmit \( Y = -1 \)
- Received signal \( \Rightarrow \) \( X = Y + Z \)
- Decoding rule \( \Rightarrow \hat{Y} = 1 \) if \( X \geq 0 \), \( \hat{Y} = -1 \) if \( X < 0 \)
  \( \Rightarrow \text{Errors:} \) ● to the left of 0 and ● to the right

\[
\hat{Y} = -1 \quad \text{and} \quad \hat{Y} = 1
\]

- Q: What is the probability of error, \( P_e := P(\hat{Y} \neq Y) \)?
From communications channel example we know

⇒ If $Y = 1$ then $X \mid Y = 1 \sim \mathcal{N}(1, \sigma^2)$. Conditional pdf is

$$f_{X \mid Y}(x \mid 1) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-1)^2}{2\sigma^2}}$$

⇒ If $Y = -1$ then $X \mid Y = -1 \sim \mathcal{N}(-1, \sigma^2)$. Conditional pdf is

$$f_{X \mid Y}(x \mid -1) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x+1)^2}{2\sigma^2}}$$
Write probability of error by conditioning on $Y = \pm 1$ (total probability)

$$
P_e = P\left(\hat{Y} \neq Y \mid Y = 1\right) P\left(Y = 1\right) + P\left(\hat{Y} \neq Y \mid Y = -1\right) P\left(Y = -1\right)
$$

$$
= P\left(\hat{Y} = -1 \mid Y = 1\right) p + P\left(\hat{Y} = 1 \mid Y = -1\right) q
$$

According to the decision rule

$$
P_e = P\left(X < 0 \mid Y = 1\right) p + P\left(X \geq 0 \mid Y = -1\right) q
$$

But $X$ given $Y$ is normally distributed, then

$$
P_e = \frac{p}{\sqrt{2\pi}\sigma} \int_{-\infty}^{0} e^{-\frac{(x-1)^2}{2\sigma^2}} dx + \frac{q}{\sqrt{2\pi}\sigma} \int_{0}^{\infty} e^{-\frac{(x+1)^2}{2\sigma^2}} dx
$$
Conditional expectation

Markov and Chebyshev’s inequalities

Convergence of random variables

Limit theorems

Conditional probabilities

Conditional expectation
Definition of conditional expectation

▶ **Def:** For continuous RVs $X$, $Y$, *conditional expectation* is

$$
E[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X \mid Y}(x \mid y) \, dx
$$

▶ **Def:** For discrete RVs $X$, $Y$, conditional expectation is

$$
E[X \mid Y = y] = \sum_x x p_{X \mid Y}(x \mid y)
$$

▶ Defined for given $y \Rightarrow E[X \mid Y = y]$ is a number

$\Rightarrow$ All possible values $y$ of $Y \Rightarrow$ random variable $E[X \mid Y]$

▶ $E[X \mid Y]$ a function of the RV $Y$, hence itself a RV

$\Rightarrow E[X \mid Y = y]$ value associated with outcome $Y = y$

▶ If $X$ and $Y$ independent, then $E[X \mid Y] = E[X]$
Conditional expectation example

- Consider independent Bernoulli RVs $Y$ and $Z$, define $X = Y + Z$

- **Q:** What is $\mathbb{E} [X \mid Y = 0]$? Recall we found the conditional pmf

  \[
  p_{X \mid Y}(0 \mid 0) = 1 - p, \quad p_{X \mid Y}(1 \mid 0) = p, \quad p_{X \mid Y}(2 \mid 0) = 0 \\
  p_{X \mid Y}(0 \mid 1) = 0, \quad p_{X \mid Y}(1 \mid 1) = 1 - p, \quad p_{X \mid Y}(2 \mid 1) = p
  \]

- Use definition of conditional expectation for discrete RVs

  \[
  \mathbb{E} [X \mid Y = 0] = \sum_{x} x p_{X \mid Y}(x \mid 0) \\
  = 0 \times (1 - p) + 1 \times p + 2 \times 0 = p
  \]
Iterated expectations

- If $E[X | Y]$ is a RV, can compute expected value $E_Y[E_X[X | Y]]$
  
  Subindices clarify innermost expectation is w.r.t. $X$, outermost w.r.t. $Y$

- What is $E_Y[E_X[X | Y]]$? Not surprisingly $\Rightarrow E[X] = E_Y[E_X[X | Y]]$

- Show for discrete RVs (write integrals for continuous)

  $$E_Y[E_X[X | Y]] = \sum_y E_X[X | Y = y] p_Y(y) = \sum_y \left[ \sum_x x p_{X|Y}(x | y) \right] p_Y(y)$$
  
  $$= \sum_x x \left[ \sum_y p_{X|Y}(x | y) p_Y(y) \right] = \sum_x x \left[ \sum_y p_{XY}(x, y) \right]$$
  
  $$= \sum_x x p_X(x) = E[X]$$

- Offers a useful method to compute expected values

  $\Rightarrow$ Condition on $Y = y$
  
  $\Rightarrow$ Compute expected value over $X$ for given $y$
  
  $\Rightarrow$ Compute expected value over all values $y$ of $Y$ $\Rightarrow E_Y[E_X[X | Y]]$
Iterated expectations example

Consider a probability class in some university

⇒ Seniors get an $A = 4$ w.p. 0.5, $B = 3$ w.p. 0.5
⇒ Juniors get a $B = 3$ w.p. 0.6, $C = 2$ w.p. 0.4
⇒ An exchange student is a senior w.p. 0.7, and a junior w.p. 0.3

Q: Expectation of $X =$ exchange student’s grade?

Start by conditioning on standing

$$
\mathbb{E}[X \mid \text{Senior}] = 0.5 \times 4 + 0.5 \times 3 = 3.5
$$
$$
\mathbb{E}[X \mid \text{Junior}] = 0.6 \times 3 + 0.4 \times 2 = 2.6
$$

Now sum over standing’s probability

$$
\mathbb{E}[X] = \mathbb{E}[X \mid \text{Senior}] P(\text{Senior}) + \mathbb{E}[X \mid \text{Junior}] P(\text{Junior})
= 3.5 \times 0.7 + 2.6 \times 0.3 = 3.23
$$
Conditioning on sum of Poisson RVs

- Consider independent Poisson RVs $Y$ and $Z$, parameters $\lambda_1$ and $\lambda_2$
- Define $X = Y + Z$. What is $\mathbb{E}[Y \mid X = x]$?
  - We found $Y \mid X = x$ is binomial $(x, \lambda_1/(\lambda_1 + \lambda_2))$, hence
  
  \[
  \mathbb{E}[Y \mid X = x] = \frac{x\lambda_1}{\lambda_1 + \lambda_2}
  \]

- Now use iterated expectations to obtain $\mathbb{E}[Y]$  
  - Recall $X$ is Poisson with parameter $\lambda = \lambda_1 + \lambda_2$
  
  \[
  \mathbb{E}[Y] = \sum_{x=0}^{\infty} \mathbb{E}[Y \mid X = x] p_X(x) = \sum_{x=0}^{\infty} \frac{x\lambda_1}{\lambda_1 + \lambda_2} p_X(x) = \frac{\lambda_1}{\lambda_1 + \lambda_2} \mathbb{E}[X] = \frac{\lambda_1}{\lambda_1 + \lambda_2} (\lambda_1 + \lambda_2) = \lambda_1
  \]

- Of course, since $Y$ is Poisson with parameter $\lambda_1$
Conditioning to compute expectations

- As with probabilities, conditioning is useful to compute expectations
  - Spreads difficulty into simpler problems (divide and conquer)

**Example**

- A baseball player scores $X_i$ runs per game
  - Expected runs are $E[X_i] = E[X]$ independently of game

- Player plays $N$ games in the season. $N$ is random (playoffs, injuries?)
  - Expected value of number of games is $E[N]$

- What is the expected number of runs in the season? $\Rightarrow E\left[\sum_{i=1}^{N} X_i\right]$

- Both $N$ and $X_i$ are random, and here also assumed independent
  - The sum $\sum_{i=1}^{N} X_i$ is known as compound RV
Sum of random number of random quantities

Step 1: Condition on $N = n$ then

$$\left[ \sum_{i=1}^{N} X_i \mid N = n \right] = \sum_{i=1}^{n} X_i$$

Step 2: Compute expected value w.r.t. $X_i$, use $N$ and the $X_i$ independent

$$\mathbb{E}_{X_i} \left[ \sum_{i=1}^{N} X_i \mid N = n \right] = \mathbb{E} \left[ \sum_{i=1}^{n} X_i \right] = n \mathbb{E} [X]$$

$\Rightarrow$ Second equality possible because $n$ is a number (not a RV)

Step 3: Compute expected value w.r.t. values $n$ of $N$

$$\mathbb{E}_N \left[ \mathbb{E}_{X_i} \left[ \sum_{i=1}^{N} X_i \mid N \right] \right] = \mathbb{E}_N \left[ N \mathbb{E} [X] \right] = \mathbb{E} [N] \mathbb{E} [X]$$

Yielding result $\Rightarrow \mathbb{E} \left[ \sum_{i=1}^{N} X_i \right] = \mathbb{E} [N] \mathbb{E} [X]$
Ex: Suppose $X$ is a geometric RV with parameter $p$

- Calculate $E[X]$ by conditioning on $Y = 1 \{ \text{“first trial is a success”} \}$
  - If $Y = 1$, then clearly $E[X \mid Y = 1] = 1$
  - If $Y = 0$, independence of trials yields $E[X \mid Y = 0] = 1 + E[X]$ 

- Use iterated expectations

\[
E[X] = E[X \mid Y = 1] P(Y = 1) + E[X \mid Y = 0] P(Y = 0)
\]
\[
= 1 \times p + (1 + E[X]) \times (1 - p)
\]

- Solving for $E[X]$ yields

\[
E[X] = \frac{1}{p}
\]

- Here, direct approach is straightforward (geometric series, derivative)
  - Oftentimes simplifications can be major
The trapped miner example

- A miner is trapped in a mine containing three doors
- At all times $n \geq 1$ while still trapped
  - The miner chooses a door $D_n = j, j = 1, 2, 3$
  - Choice of door $D_n$ made independently of prior choices
  - Equally likely to pick either door, i.e., $P(D_n = j) = 1/3$
- Each door leads to a tunnel, but only one leads to safety
  - Door 1: the miner reaches safety after two hours of travel
  - Door 2: the miner returns back after three hours of travel
  - Door 3: the miner returns back after five hours of travel

- Let $X$ denote the total time traveled till the miner reaches safety
- **Q:** What is $\mathbb{E}[X]$?
The trapped miner example (continued)

- Calculate \( \mathbb{E}[X] \) by conditioning on first door choice \( D_1 \)
  - If \( D_1 = 1 \), then 2 hours and out, i.e., \( \mathbb{E}[X | D_1 = 1] = 2 \)
  - If \( D_1 = 2 \), door choices independent so \( \mathbb{E}[X | D_1 = 2] = 3 + \mathbb{E}[X] \)
  - Likewise for \( D_1 = 3 \), we have \( \mathbb{E}[X | D_1 = 3] = 5 + \mathbb{E}[X] \)

- Use iterated expectations

\[
\mathbb{E}[X] = \sum_{j=1}^{3} \mathbb{E}[X | D_1 = j] \cdot P(D_1 = j) = \frac{1}{3} \sum_{j=1}^{3} \mathbb{E}[X | D_1 = j]
\]

\[
= \frac{2 + 3 + \mathbb{E}[X] + 5 + \mathbb{E}[X]}{3} = \frac{10 + 2\mathbb{E}[X]}{3}
\]

- Solving for \( \mathbb{E}[X] \) yields

\[
\mathbb{E}[X] = 10
\]

- You will solve it again using compound RVs in the homework
Conditional variance formula

- **Def:** The conditional variance of $X$ given $Y = y$ is

\[
\text{var}[X|Y = y] = \mathbb{E} \left[ (X - \mathbb{E}[X|Y = y])^2 \mid Y = y \right] \\
= \mathbb{E}[X^2 \mid Y = y] - (\mathbb{E}[X \mid Y = y])^2
\]

⇒ $\text{var}[X|Y]$ a function of RV $Y$, value for $Y = y$ is $\text{var}[X|Y = y]$

- Calculate $\text{var}[X]$ by conditioning on $Y = y$. Quick guesses?

  ⇒ $\text{var}[X] \neq \mathbb{E}_Y[\text{var}_X(X \mid Y)]$

  ⇒ $\text{var}[X] \neq \text{var}_Y[\mathbb{E}_X(X \mid Y)]$

- **Neither.** Following conditional variance formula is the correct way

\[
\text{var}[X] = \mathbb{E}_Y[\text{var}_X(X \mid Y)] + \text{var}_Y[\mathbb{E}_X(X \mid Y)]
\]
Proof.

- Start from the first summand, use linearity, iterated expectations

\[
\mathbb{E}_Y[\text{var}_X(X \mid Y)] = \mathbb{E}_Y \left[ \mathbb{E}_X(X^2 \mid Y) - (\mathbb{E}_X(X \mid Y))^2 \right]
\]

\[
= \mathbb{E}_Y \left[ \mathbb{E}_X(X^2 \mid Y) \right] - \mathbb{E}_Y \left[ (\mathbb{E}_X(X \mid Y))^2 \right]
\]

\[
= \mathbb{E} [X^2] - \mathbb{E}_Y \left[ (\mathbb{E}_X(X \mid Y))^2 \right]
\]

- For the second term use variance definition, iterated expectations

\[
\text{var}_Y[\mathbb{E}_X(X \mid Y)] = \mathbb{E}_Y \left[ (\mathbb{E}_X(X \mid Y))^2 \right] - (\mathbb{E}_Y[\mathbb{E}_X(X \mid Y)])^2
\]

\[
= \mathbb{E}_Y \left[ (\mathbb{E}_X(X \mid Y))^2 \right] - (\mathbb{E} [X])^2
\]

- Summing up both terms yields (blue terms cancel)

\[
\mathbb{E}_Y[\text{var}_X(X \mid Y)] + \text{var}_Y[\mathbb{E}_X(X \mid Y)] = \mathbb{E} [X^2] - (\mathbb{E} [X])^2 = \text{var} [X]
\]
Let $X_1, X_2, \ldots$ be i.i.d. RVs with $E[X_1] = \mu$ and $\text{var}[X_1] = \sigma^2$.

Let $N$ be a nonnegative integer-valued RV independent of the $X_i$.

Consider the compound RV $S = \sum_{i=1}^{N} X_i$. What is $\text{var}[S]$?

The conditional variance formula is useful here.

Earlier, we found $E[S|N] = N\mu$. What about $\text{var}[S|N = n]$?

$$\text{var}\left[\sum_{i=1}^{N} X_i|N = n\right] = \text{var}\left[\sum_{i=1}^{n} X_i|N = n\right] = \text{var}\left[\sum_{i=1}^{n} X_i\right] = n\sigma^2$$

$$\Rightarrow \text{var}[S|N] = N\sigma^2.$$ Used independence of $N$ and the i.i.d. $X_i$.

The conditional variance formula is $\text{var}[S] = E[N\sigma^2] + \text{var}[N\mu]$.

Yielding result $$\Rightarrow \text{var}\left[\sum_{i=1}^{N} X_i\right] = E[N]\sigma^2 + \text{var}[N]\mu^2.$$
Glossary

- Markov’s inequality
- Chebyshev’s inequality
- Limit of a sequence
- Almost sure convergence
- Convergence in probability
- Mean-square convergence
- Convergence in distribution
- I.i.d. random variables
- Sample average
- Centering and scaling
- Law of large numbers
- Central limit theorem
- Conditional distribution
- Communication channel
- Probability of error
- Conditional expectation
- Iterated expectations
- Expectations by conditioning
- Compound random variable
- Conditional variance