First a note about fonts, math symbols and errors in these slides

• I do not use a math package such as MathType, a LaTeX plugin, or the native equation editor, just ascii/unicode with only "first order" superscripts and/or subscripts
• I use the following unicode math symbols:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>∫</td>
<td>Integral</td>
</tr>
<tr>
<td>∬</td>
<td>Double integral</td>
</tr>
<tr>
<td>∭</td>
<td>Triple integral</td>
</tr>
<tr>
<td>∑</td>
<td>Summation</td>
</tr>
<tr>
<td>∏</td>
<td>Product</td>
</tr>
<tr>
<td>∃</td>
<td>There exists</td>
</tr>
<tr>
<td>∄</td>
<td>There does not exist</td>
</tr>
<tr>
<td>∀</td>
<td>For all</td>
</tr>
<tr>
<td>∈</td>
<td>Element of</td>
</tr>
<tr>
<td>∉</td>
<td>Not element of</td>
</tr>
<tr>
<td>∋</td>
<td>Contains as element</td>
</tr>
<tr>
<td>∈</td>
<td>Does not contain as element</td>
</tr>
<tr>
<td>∖</td>
<td>Set minus</td>
</tr>
<tr>
<td>⊇</td>
<td>Superset</td>
</tr>
<tr>
<td>⊆</td>
<td>Subset</td>
</tr>
<tr>
<td>⊂</td>
<td>Proper subset</td>
</tr>
<tr>
<td>⊃</td>
<td>Proper superset</td>
</tr>
<tr>
<td>⊄</td>
<td>Not subset</td>
</tr>
<tr>
<td>⊅</td>
<td>Not superset</td>
</tr>
<tr>
<td>⊆</td>
<td>Subset</td>
</tr>
<tr>
<td>⊇</td>
<td>Superset</td>
</tr>
<tr>
<td>≠</td>
<td>Not equal</td>
</tr>
<tr>
<td>≈</td>
<td>Approximately equal</td>
</tr>
<tr>
<td>≠</td>
<td>Not equal</td>
</tr>
<tr>
<td>≡</td>
<td>Equal</td>
</tr>
<tr>
<td>≢</td>
<td>Not equal</td>
</tr>
<tr>
<td>≤</td>
<td>Less than or equal</td>
</tr>
<tr>
<td>≥</td>
<td>Greater than or equal</td>
</tr>
<tr>
<td>≪</td>
<td>Much less than</td>
</tr>
<tr>
<td>≫</td>
<td>Much greater than</td>
</tr>
<tr>
<td>∴</td>
<td>Therefore</td>
</tr>
<tr>
<td>∵</td>
<td>Because of</td>
</tr>
<tr>
<td>⊕</td>
<td>Direct sum</td>
</tr>
<tr>
<td>⊂</td>
<td>Proper subset</td>
</tr>
<tr>
<td>⊃</td>
<td>Proper superset</td>
</tr>
<tr>
<td>∈</td>
<td>Element of</td>
</tr>
<tr>
<td>∉</td>
<td>Not element of</td>
</tr>
<tr>
<td>∋</td>
<td>Contains as element</td>
</tr>
<tr>
<td>∈</td>
<td>Does not contain as element</td>
</tr>
<tr>
<td>∖</td>
<td>Set minus</td>
</tr>
<tr>
<td>⊇</td>
<td>Superset</td>
</tr>
<tr>
<td>⊆</td>
<td>Subset</td>
</tr>
<tr>
<td>⊂</td>
<td>Proper subset</td>
</tr>
<tr>
<td>⊃</td>
<td>Proper superset</td>
</tr>
<tr>
<td>⊄</td>
<td>Not subset</td>
</tr>
<tr>
<td>⊅</td>
<td>Not superset</td>
</tr>
<tr>
<td>⊆</td>
<td>Subset</td>
</tr>
<tr>
<td>⊇</td>
<td>Superset</td>
</tr>
<tr>
<td>∴</td>
<td>Therefore</td>
</tr>
<tr>
<td>∵</td>
<td>Because of</td>
</tr>
<tr>
<td>⊕</td>
<td>Direct sum</td>
</tr>
</tbody>
</table>

- ∃ \( \frac{\pi}{2} \), \( \infty \), \( \land \), \( \lor \), \( \in \), \( \notin \), \( \subseteq \), \( \supseteq \), \( \subseteq \), \( \supseteq \), \( \leq \), \( \geq \)

So some things won't look nice (such as summation and fractional displays) and there are some notational peculiarities because I try to avoid sub-sub-scripting and the like.

• For example, \( X_1 \) as a subscript will appear as \( X_1 \), i.e., \( F_{X_1} \)
• Please email me at kesidis@gmail.com with any corrections and suggestions.
Outline of the course

- Introduction to probability and counting
- Random variables (RVs) and their distributions
- Conditioning and independence, sums of independent RVs
- Useful inequalities and the weak law of large numbers
- Central limit theorem and statistical confidence
- Wide-sense stationary processes and linear, time-invariant filtering.

References

Introduction to probability – outline

• Get our feet dirty by tossing a pair of dice to illustrate naïve probability theory.
• In the process, introduce a bunch of terms that we will later more carefully define.
• Introduction to Combinatorics (counting things).
• Axiomatic probability theory based on set and measure theory.
• De Morgan's theorem as a vehicle to discuss methods of proof.
• Independent and conditional events.
• Tree diagrams.
Uncertainty in engineering

• Dealing with uncertainty is fundamental to engineering.
• A well-engineered system operating in a complex environment has carefully delineated what is known and what is uncertain.
• What is uncertain may need to be continually reassessed as, e.g., as information is observed eliminating some uncertainty, or as the system moves to a new environment creating some uncertainty.
• Through experience (statistics), one can obtain a mathematical model describing the laws governing uncertainty – this is the role of probability theory.

Random experiment throwing two dice

• Historically, naïve probability arose through the consideration of random experiments with equally likely outcomes (samples).
• For example, consider a random experiment consisting of tossing two standard dice each with 6 sides (white cubes) enumerated 1-6 (number of black dots on each side).
• Suppose the outcomes are the numbers on the upturned faces of the dice after they come to rest, e.g., (2,3) is the outcome when the first die is 2 and the second die is 3.
• The set of outcomes/samples is the sample space of the random experiment, that here can be characterized as the Cartesian set product
  \[ \{1,2,3,4,5,6\} \times \{1,2,3,4,5,6\} = \{(n,m) | n,m \in \{1,2,3,4,5,6\}\} \equiv \Omega \]
• The total number of possible outcomes is \(|\Omega| = 6 \times 6 = 36\).
• Assume that such each outcome is equally likely (i.e., the dice are fair), so that the (nonnegative) probability of a given outcome is the inverse of the total number of possible outcomes, i.e., \(1/|\Omega| = 1/36\).
• So, the sum of the probabilities of all samples = 1.
Sum of two dice

- The game of craps consists of a sequences of independent tosses of a pair of putatively fair dice (i.e., independent trials of this random experiment).
- In craps, one is interested in the sum of the upturned faces of the dice.
- To study this game, we could redefine the sample space accordingly to be \( \Omega' = \{2,3,4,...,12\} \), but now each sample is not equally likely (next slides).
- Alternatively, we can keep the original sample space \( \Omega \) and define a random variable mapping \( \Omega \) to the set of real numbers,
  
  \[ X: \Omega \rightarrow \mathbb{R}, \]

  so that
  
  \[ X(n,m) = n+m. \]
- The following table depicts this random variable.

<table>
<thead>
<tr>
<th>2nd die</th>
<th>1st die</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>5</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>6</td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>7</td>
<td>8</td>
<td>9</td>
<td>10</td>
<td>11</td>
<td>12</td>
</tr>
</tbody>
</table>
Sum of two dice (cont.)

- Note that $X=7$ for six different samples/outcomes: $(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)$, i.e., we've identified the event $\{X=7\} = \{s \in \Omega \mid X(s)=7\} = \{(1,6), (2,5), (3,4), (4,3), (5,2), (6,1)\} \subseteq \Omega$
- So, the probability that the event $X=7$ occurs (the probability that the random experiment produces an outcome in this event) is simply the size of the event divided by the size of the sample space $P(X=7) = |\{X=7\}| / |\Omega| = 6 / 36 = 1/6$
- I.e., $P(X=7)$ is the sum of the probabilities of the samples in $\{X=7\}$.
- We will later obtain similar results by counting for experiments involving coin tossing and cards drawn from a deck, especially stud poker hands.
- Note that we have arrived at this result by counting outcomes, not experimentally by actually tossing a pair of fair dice.

Probability through statistics

- Instead of presuming the dice were fair, we can attempt to discover the underlying probability law of random experiment that involves tossing them by simply conducting independent trials of this random experiment.
- That is, suppose we toss the dice $n$ times and observe how many times the sum is 7, $T_7(n)$.
- If the dice were fair, one would expect that the time/trial average $T_7(n)/n \rightarrow 1/6$ as $n \rightarrow \infty$.
- This is an example of a law of large numbers (or an ergodic property), relating a statistic $T_7(n)/n$ of $n$ experiments to a putative underlying probability law governing them, $P(X=7) = 1/6$. 
Probability through statistics – random.org

- Using atmospheric noise, random (i.e., equally likely, uniformly random) and independent samples from \( \Omega = \{1, 2, 3, ..., 100\} \) are tabulated at the website random.org.

- Let \( T_k(n) \) be how many times the number \( k \) appears in the first \( n \) samples of the random.org table.
  
  \[
  T_k(n) = \sum_{i=1}^{n} 1\{X(i) = k\},
  \]
  
  where \( X(i) \) is the \( i^{th} \) sample, and \( 1A = 1 \) when \( A \) occurs & otherwise \( = 0 \).

- The following two graphs plot a single sample path of \( T_5(n)/n \) versus \( n \) and \( T_{23}(n)/n \) versus \( n \), respectively, each of which are seen to be converging to 0.01 = 1/100.

- Later, to prove the weak law of large numbers, we will establish that \( T_k(n)/n \) is a random variable with
  
  - mean (average value, expectation) \( E T_k(n)/n = 1/100 \),
  
  - variance → 0 as \( n \to \infty \), and

  - sample space \( \Omega = \{1, 2, ..., 100\} \), a product space where each sample is a vector of \( n \) samples from the table.

- Also, we will understand \( T_7(n)/n \) or \( T_7(n) \) as discrete-time stochastic processes in \( n \).

T_5(n)/n sample path from random.org
Sum of two dice – De Mere’s paradox

• Consider the following argument for computing the P(X=6):
  – There are three ways this occurs, (1,5), (2,4) and (3,3)
  – There are 21 possible outcomes, (1,1),(1,2),(1,3),(1,4),(1,5),(1,6),
    (2,2),(2,3),(2,4),(2,5),(2,6), (3,3), (3,4),(3,5),(3,6),
    (4,4),(4,5),(4,6), (5,5), (5,6), (6,6)
  – So, P(X=6) = 3/21 = 1/7
• However, if we conduct repeated trials with fair dice, we find that
  $T_6(n)/n \rightarrow 5/36$ as $n \rightarrow \infty$.
• The problem is that we have not counted outcomes properly in the
  argument above (as Pascal originally explained).
• Though there is only one way that an outcome (3,3) can occur, each die
  gives 3, there are two ways an outcome (1,5) can occur: one die gives 1
  and the other 5 and vice versa.
• So, we have undercounted the number of ways X=6 (3 instead of 5)
  and the number of outcomes in the sample space (21 instead of 36).
Sum of two dice – De Mere’s paradox (cont)

- The incorrect argument at the top of this slide counts combinations while the correct one in earlier slides counts permutations (of the outcomes of the two different dice).
- De Mere’s paradox (mistake) is not unreasonable considering the dice themselves are usually indistinguishable to the human observer, i.e., they are typically both white cubes, of the same size and weight, and with the same black-dot patterns on their sides.

Combinatorics

- Again consider a finite space with equally likely “atomic events” (or “singleton” events with a single sample element).
- So in order to compute probability of any event A, we can count the number of samples they contain, i.e., \( P(A) = \frac{|A|}{|\Omega|} \).
- However, counting may be far from simple.
Permutations with different objects
without replacement

- Consider a random experiment wherein we draw m balls (without replacement) from an urn containing \( n \geq m \) different numbered balls and place them in a row.
- So, every outcome of this experiment is an ordered sequence of \( m \) different balls, i.e., an \( m \)-permutation of \( n \) different balls.
- So, 5,17,8 and 17,5,8 are two different 3-permutations of 20 balls numbered 1 to 20, even though they are the same 3-combination of balls.
- The number of different \( m \)-permutations (i.e., the size of the sample space) is \( n(n-1)(n-2)...(n-m+1) = n!/(n-m)! \)
- Note that there are \( n \) choices for the first ball, then \( n-1 \) choices for the second (because first ball was not replaced before the second one was chosen), and so forth; also \( n-[n-m+1]+1=m \).
- For example, the number of permutations of 5 cards drawn from a standard deck of 52 different ones is
\[
52! / 47! = 52 \times 51 \times 50 \times 49 \times 48
\]

Permutations with different objects – choice with replacement

- For choice with replacement (so that the same ball can be chosen more than once), by the same argument the number of \( m \)-permutations is simply \( n^m \).
- For example, there are \( 5^{26} \) 5-letter words that can be formed from an alphabet of 26 letters – letter order matters when forming words!
Permutations with identical objects – complete choice without replacement

- The number of n-permutations of n different objects (without replacement) is simply n factorial, \( n! = n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1 \)
- Suppose that these n objects are not all different but that there \( k < n \) types of objects with: \( m_1 \) identical objects of type 1, \( m_2 \) identical objects of type 2, etc., such that \( m_1 + m_2 + \ldots + m_k = n \).
- For a given n-permutation, note that we can permute (identical) objects of the same type, wherever they lie in the n-sequence, without changing the n-permutation.
- So, the number of distinct permutations is reduced to

\[
\frac{n!}{(m_1!m_2! \ldots m_k)!},
\]

where we note that \( 0! = 1 \) by definition.
- In the particular case of all identical balls (i.e., \( k=1 \) and \( m_1 = n \)), there is obviously only \( 1 = \frac{n!}{n!} \) distinct permutation.

Permutations with identical objects – complete and incomplete cases

- For example, there are \( 5!/(2!3!) \) different 5-permutations of 5 letters two of which are A and three of which are B.
- Note that each permutation in this example is a word (letter order matters in words, otherwise god would be a dog, which is not to suggest what he/she is or isn’t, one way or the other).
- To compute the number of incomplete 4-permutations (4-letter words) for this example, we need to take two cases:
  - 4-permutations of 2As and 2Bs = \( 4!/(2!2!) = 6 \)
  - 4-permutations of 1A and 3Bs = \( 4!/(3!1!) = 4 \)
  - Total number of 4-permutations is \( 6 + 4 = 10 \)
- Note that for (complete) 5-permutations, obviously all instances involve the same number of As (2) and Bs (3).
Example:
Diagram of composition of balls in an urn

Permutations with identical objects - example

- The previous figure shows the composition of balls to be chosen at random from an urn.
- We see along the bottom that $m_1 = 2$ (black, light, even-numbered balls), $m_2 = 3$ (black, light, odd-numbered balls), etc.
- Suppose a (complete) 30-permutation is formed by randomly selecting all of the balls from the urn without replacement.
- The number of different 30-permutations, jointly considering all three features of the balls, is $30!/2!3!5!0!6!4!4!6!$ where again recall $0! = 1$ by def’n.
Permutations with identical objects – probability example

• The probability that the first 10 balls selected are black is 10!20!/30!, where the other ball-attributes were ignored here.

• To see why, suppose that a 30-permutation is the outcome of a random experiment and that all 30-permutations are equally likely.

• So the size of (number of samples in) the sample space is |Ω|=30! and the size of the event in question is |A|=10!20! so that

\[ P(A) = \frac{|A|}{|Ω|} \]

• Exercise: Why is |Ω|=30! here instead of the figure given at the bottom of the previous slide (i.e., the number of different permutations considering the 3 features of the balls)? Hint: Recall De Mere’s paradox.

Combinations

• Suppose that the order in which different items were chosen doesn’t matter; i.e., we are just interested in the number of combinations not their order (permutations).

• Recall, when choosing m objects from n different objects without replacement, the number of m-permutations of n objects is

\[ \frac{n!}{(n-m)!} \]

• Since the order of the m objects does not matter in a combination, we divide by m! to obtain the number of distinct combinations of m-objects chosen from n objects without replacement as

\[ C(n,m) = \frac{n!}{[(n-m)! \cdot m!]} \]

read “n choose m”.
Combinations: 5-card stud poker hands

- A standard deck of 52 different playing cards consists of cards of
  - 4 different suits (hearts, clubs, spades and diamonds), and
  - 13 different denominations (2,3,...,10,J,Q,K,A) for each suit, where
    aces (A) are highest rank and deuces (2) are lowest.
- So, each card has two features: denomination and suit.
- The number of different 5-card “hands” (order doesn’t matter, choice
  without replacement) is \( C(52,5) \).
- A flush is a 5-card hand consisting only of (different) cards of the same
  suit.
- If a hand is chosen uniformly at random, the probability of choosing a
  flush 5-card hand is \( 4C(13,5)/C(52,5) \) because there are 4 suits and
  \( C(13,5) \) flush combinations of each suit.
- This is not quite right as we should subtract the number of straight-
  flushes (which is 36) in the numerator, cf. the discussion of straights
  below.

Full house 5-card stud poker hands

- A full house is a hand with two cards of one denomination and three
  cards of another.
- So, the number of full house hands is
  \[ C(13,1)C(4,3) \times C(12,1)C(4,2), \]
  where \( C(4,3) \) = number of suit-combinations for the 3-of-a-kind.
- Here we first need to identify which denomination is the thee-of-a-kind
  and then which denomination is two-pair (or vice versa), i.e., order
  matters in this sense.
- E.g., Aces over fours (3 Aces and 2 fours) is different from fours over
  aces.
- For this reason, the number of different full house hands is not
  \( C(13,2) \times C(4,3)C(4,2) \).
- We could decide the pair before the three-of-a-kind, leading to the
  same answer, i.e., \( 13C(4,2) \times 12C(4,3) = 13C(4,3) \times 12C(4,2). \)
- Again, to find the probability of choosing a full house, divide this
  number of full-house hands by \( C(52,5) \).
Two-pair 5-card stud-poker hands

• Two pair is a hand with two cards of one denomination and two other cards of another denomination (to avoid four-of-a-kind), and the 5th, non-pair card is of a third denomination (so not a full house).
• Here, order does not matter in the choice of pair denominations, but it does matter in the choice of pair denominations and the non-pair card.
• So, the number of two-pair hands is \(C(13,2)C(4,2)C(4,2) \times 44\) where the last term is the 5th card in the hand (again, one of 44 = \(C(11,1)C(4,1)\) cards from a denomination different from those of the two-pair).
• Note that we could have chosen the non-pair card first and arrived at the same answer: \(C(13,1)C(4,1)C(12,2)C(4,2)^2\)

• Exercise: Check that the number of full houses is less than then number of flushes which is less than the number of two-pair hands; so a full house ranks higher than a flush which ranks higher than two pair.
• Exercise: Argue that the number of hands that are one-pair is \(C(13,1)C(4,2)C(12,3)C(4,1)^2 = C(13,3)C(4,1)^3C(10,1)C(4,2)\)
• Exercise: Argue that the number of hands that are three-of-a-kind is \(C(13,1)C(4,3)C(12,2)C(4,1)^2 = C(13,2)C(4,1)^2C(11,1)C(4,3)\)

Combinations – straights in stud poker hands

• A straight is a hand consisting of five (denominationally) consecutive cards
• There are 9 = 10 - 2 + 1 different types of straights say based on the denomination of the lowest card \(\{2,3,...,10\}\) (again, with aces high).
• So, by accounting for the suits, the number of different straight hands is \(9 \times C(4,1)^5 = 9 \times 4^5\) (order matters when counting straights).
• This answer is actually not quite right as it includes straight-flushes (hands that are both straights and flushes), the highest ranked hands in 5-card stud poker.
• The number of straight-flushes is \(9 \times 4 = 36\).
• So, the correct number of straights is \(9 \times 4^5 – 36\).
• Exercise: Where should straights rank among the hands described so far?
• Exercise: Show the number of four-of-a-kind hands is \(C(13,1)C(4,4)C(12,1)C(4,1)\) and argue why this hand should rank less than a straight-flush.
Combinations – different-card poker hands

- **Exercise:** Explain why the number of different lowest ranked hands in poker, 5 cards all of different denominations neither forming a straight nor a flush, is

\[ C(13,5) \times 4^5 - 9 \times 4^5 - 4C(13,5) + 9 \times 4 \]

Two poker hands simultaneously drawn

- If two players are each dealt 5 cards from the same 52-card deck (without replacement), what is the probability that they get equally ranked straights?
- To answer: First, the number of different pairs of 5-card hands that can be drawn is \( C(52,10)C(10,5) = C(52,5)C(47,5) \)
- Now compute the number of ways of drawing two equally ranked straights.
- There are 14-6+1=9 different ranks for straights based on the card with highest denomination (Ace, King, Queen, Jack, 10, 9, 8, 7, 6)
- Given this, the denominations of all cards are set in a straight, so what remains is to choose the suits: \( (C(4,2)C(2,1))^5 \)
- So, the desired probability is: \( 9(C(4,2)C(2,1))^5/(C(52,10)C(10,5)) \)
- **Exercise:** Repeat this for 3 players and for tied hands that are two-pair, flushes or straight-flushes.
- **Computer Exercise:** When playing against one other player, compute the probability that your hand will beat your opponent’s, given whichever particular hand is yours.
Farkle

• Consider a fair throw of six 6-sided dice.
• The probability of a straight is $1/5!$
• Probability of 6 of a kind is $C(6,1)/6^6 = 1/6^5$ where $C(6,1)=6$ is the number of denominations of the 6 of a kind.
• The probability of 2 triplets is
  \[ C(6,2)C(6,3)C(3,3)/6^6 = C(6,2)C(6,3)/6^6 \]
  where
  – $C(6,2)$ chooses the denominations of the 2 triplets (e.g., ones and fives),
  – $C(6,3)$ chooses the dice for the first denomination, and
  – $C(3,3)=1$ chooses the remaining dice for the other denomination.

Farkle (cont)

• There are two ways for three pair: three different denominations of pairs and one "full house" (four of a kind and one denominationally different pair).
• In the former case, we choose the denominations as $C(6,3)$, as they are all pairs and so choice-order doesn’t matter.
• In the latter case, we choose denominations as a permutation, $C(6,1)C(5,1)=6\cdot5$, because order does matter when designating the four-of-a-kind versus the (single) pair.
• So, the probability of three-pair is
  \[ [C(6,3)C(6,2)C(4,2)C(2,2) + C(6,1)C(6,4)C(5,1)C(2,2)]/6^6 \]
• The probability of just one pair is $C(6,1)C(6,2)C(6,4)4!/6^6$ where
  – $C(6,2)$ chooses the dice for the pair, and
  – $4!$ chooses the unpaired dice (once their denominations are set in $C(6,4)$ ways).
Combinations for choice with replacement

Consider an urn with \( n \) different balls and suppose we are interested in the number of combinations of choices of \( m \) balls where every chosen ball is replaced in the urn.

So, under choice with replacement, a given ball can appear up to \( m \)-times in an \( m \)-combination.

To compute the number of combinations, we need a clever encoding of a given \( m \)-combination.

Combinations with replacement (cont)

To this end, arbitrarily enumerate the \( n \) different balls, i.e., so we can identify a first ball, a second ball, etc.

The desired encoding of an \( m \)-combination of \( n \) objects chosen with replacement is an ordering of \( n+1 \) symbols consisting of \( m \) identical symbols “|” representing choice of the object on the left and \( n-1 \) identical symbols “o” representing balls numbered 2,3,...,\( n \).

For example, for \( n=5 \) and \( m=2 \),
- \( o\ o\ o\ |\ o \) represents the choice of ball 3 and ball 4
- \( |\ o\ o\ o\ o \) represents the choice of ball 1 and ball 1
- \( o\ o\ o\ o\ | \) represents the choice of ball 5 and ball 5

So, the number of unique such permutations, accounting for the identical symbols, is

\[
(\text{n-1+m)}/[m!(\text{n-1})!] = C(n-1+m,m) = C(n-1+m,n-1)
\]
Combinations with replacement - Discussion

• Note how addition of a right-most symbol “o” to explicitly represent ball 5 would not change the number of permutations as that symbol would never “move” (it’s implicitly there as the choice symbol “|” needs an object on the left to refer to).
• Note also that, owing to choice with replacement, it’s possible that m>n.
• In summary, every combination-with-replacement is encoded as an ordered list of two types of symbols; so the number of combinations-with-replacement is equal to the number of permutations of this encoding.

Combinations for choice with replacement - example

• Suppose we choose (uniformly at random) a five-card hand from a standard deck of 52 playing cards, but with replacement.
• Find the probability of obtaining five aces.
  – The total number of 5-card hands under choice with replacement is \( C(52-1+5,5) \)
  – To find the number of hands with 5 aces under choice with replacement is \( C(4-1+5,5) \), i.e., there are 4 different aces in the deck.
  – So, the probability of choosing 5 aces is \( C(8,5) / C(56,5) \).
• Note that if order matters, the probability of 5-ace sequence for choice with replacement would be \( 4^5 / 52^5 \).
Example: ways to make a sum

- How many different permutations of $k$ non-negative integers sum to $n$?
- For this problem, we have $n$ identical objects “1” and $k-1$ identical objects “|” which we are permuting.
- Let $x_1 \geq 0$ be the number of 1’s to the left of the leftmost (first) |
- For $j=2,3,k-1$, let $x_j \geq 0$ be the number of 1’s between $(j-1)^{th}$ | and $j^{th}$ |
- Let $x_k \geq 0$ be the number of 1’s to the right of the $(k-1)^{th}$ |
- So, the answer is $C(n+k-1,n)=C(n+k-1,k-1)$.
- For $n=7$ and $k=3$, |111|1111 corresponds to $(x_1,x_2,x_3)=(0,3,4)$, and $C(7+3-1,7)=36$.
- **Exercise:** Simply adapt this argument for the case of strictly positive $x_i>0$.

Placing distinguishable objects into indistinguishable nonempty bins

- $S(n,k)$ = number of ways to place distinguishable objects into indistinguishable, nonempty bins,
- i.e., each bin is given at least one object so that $n \geq k$ is required.
- First note that $\forall n \geq 1$, $S(n,n)=1=S(n,1)$.
- The following recursion is now explained:
- $\forall n \geq k \geq 1$, $S(n+1,k) = k \cdot S(n,k) + S(n,k-1)$
- Consider two cases for placement of the $(n+1)^{st}$ object:
  - It’s placed alone in a bin, hence the other $n$ objects are placed in the remaining $k-1$ bins, done in $S(n,k-1)$ ways
  - It’s placed in one of $k$ non-empty bins, done in $k \cdot S(k,n)$ ways
Placing distinguishable objects into indistinguishable nonempty bins

- **Exercise:** Prove by induction that
  \[ \forall n \geq k \geq 1, \ S(n,k) = (k!)^{-1} \sum_{j=0}^{k} (-1)^{k-j} \binom{k}{j} j^n \]
- **Exercise:** Argue by inclusion-exclusion for this expression for \( S(n,k) \).
- **Exercise:** What if the bins are distinct? That is, how many ways to place \( n \) distinguishable objects into \( k \) distinguishable bins.
- **Exercise:** What if \( k \) distinguishable objects into \( n \) distinct bins so that exactly \( m \leq n \) are not empty?
- \( S(n,k) \) are called Stirling numbers of the second kind.

Examples from elementary statistical mechanics

- We now give some applications of combinatorics and naïve probability theory to elementary statistical mechanics.
- To that end, we will first give some background on constrained optimization.
- Some of the concepts touched on in these examples (e.g., multinomial distributions and Stirling’s approximation) subsequently will be covered in greater detail.
Background on constrained optimization and duality

• Consider a primal optimization problem with a set of $m$ inequality constraints: Find 
  \[ \arg\min_{x \in D} f_0(x), \]
  where the constrained domain of optimization is 
  \[ D \equiv \{ x \in \mathbb{R}^n \mid f_i(x) \leq 0, \forall i \in \{1,2,\ldots,m\} \}. \]
• To study the primal problem, we define the corresponding Lagrangian function on $\mathbb{R}^{n+m}$:
  \[ L(x, v) \equiv f_0(x) + \sum_{i=1}^{m} v_i f_i(x) \]
  where, by implication, the vector of Lagrange multipliers is 
  \[ v \in \mathbb{R}^m. \]
• Let $v \geq 0$ connote that all of the $m$ Lagrange multipliers are nonnegative, i.e., $v \in [0, \infty)^m$.

Primal constrained optimization with inequality constraints - Lagrange multipliers

• **Theorem:**
  \[ \min_{x \in \mathbb{R}^n} \max_{v \geq 0} L(x, v) = \min_{x \in D} f_0(x) \equiv p*. \]
• **Proof:** Simply,
  if $x \notin D$ then $\max_{v \geq 0} L(x, v) = \infty$,
  else $\max_{v \geq 0} L(x, v) = f_0(x)$. Q.E.D.
• So, we can maximize the Lagrangian in an unconstrained fashion to find the solution to the constrained primal problem.
Complementary slackness of the primal solution

- Define the maximizing values of the Lagrange multipliers,
  \[ v^*(x) \equiv \arg \max_{v \geq 0} L(x, v) \]
  and note that the complementary slackness conditions
  \[ v_i^*(x)f_i(x) = 0 \]
  hold for all \( x \in D \) and \( i \in \{1, 2, \ldots, m\} \).
- That is, if there is slackness in the \( i \)th constraint, i.e., \( f_i(x) < 0 \),
  then there is no slackness in the constraint of the corresponding Lagrange multiplier,
  i.e., \( v_i^*(x) = 0 \).
- Conversely, if \( f_i(x) = 0 \), then the optimal value of the Lagrange multiplier \( v_i^*(x) \)
  is not relevant to the Lagrangian.
- Complementary slackness conditions lead to the Karush-Kuhn-Tucker necessary conditions
  for optimality of the primal solution.

The dual problem

- Now define the dual function of the primal problem:
  \[ G(v) := \min_{x \in \mathbb{R}^n} L(x, v). \]
- Note that \( G(v) \) may be infinite for some values of \( v \) and that \( g \) is always concave.
- **Theorem:** For all \( x \in D \) and \( v \geq 0 \),
  \[ G(v) \leq f_0(x). \]
- **Proof:** For all \( v \geq 0 \),
  \[ G(v) \leq L(x, v) \leq \max_{v \geq 0} L(x, v) = f_0(x), \]
  where the last equality is the bound on \( L \) assuming \( x \in D \).
The dual problem (cont)

- So, by the previous theorem, if we solve the dual problem, i.e., find
  \[ d^* \equiv \max_{v \geq 0} G(v), \]
  then we will have obtained a (hopefully good) lower bound to the primal problem, i.e.,
  \[ d^* \leq p^*. \]

The duality gap

- Under certain conditions in this finite dimensional setting, in particular when the primal problem is convex and a strictly feasible solution exists, the duality gap
  \[ p^* - d^* = 0. \]
- To use duality to find \( p^* \) and \( \arg \max_{x \in D} f_0(x) \) in this case, suppose that a slow ascent method is used to numerically maximize \( g \),
  \[ v_n = v_{n-1} + \alpha_1 \nabla G(v_{n-1}), \]
  and between steps of the slow ascent method, a fast descent method is used to evaluate \( G(v_n) \) by minimizing \( L(x, v_n) \),
  \[ x_k = x_{k-1} - \alpha_2 \nabla_x L(x_{k-1}, v_n). \]
- The process described by such an ascent/descent method is called an iterative sub-gradient method.
- Note that the step size \( \alpha_1 \) may need to be chosen so that
  \[ v_n \geq 0 \] for all \( n \).
Equality constraints

- If $f_0$ is convex, then we can solve the dual problem
  \[ \bar{x}^*(\nu) := \arg \min_{x \in \mathbb{R}^n} L(x, \nu). \]
- Assuming that we’re dealing with equality constraints defining the domain $D$, we can then attempt to select the Lagrange multipliers $\nu^*$ so that all of the equality constraints are met:
  \[ f_j(\bar{x}^*(\nu^*)) = 0 \text{ for all } j. \]
- If such a set of Lagrange multipliers $\nu^*$ exists, then
  \[ \bar{x}^*(\nu^*) = \arg \min_{x \in D} f_0(x). \]

Some elementary applications to statistical mechanics - preliminaries

- Suppose $N$ particles independently occupy one of $J$ energy states.
- If the particles are distinguishable, the number of ways that there are $n_j$ particles in state $j$ is
  \[ C(N, n_1)C(N-n_1, n_2) \ldots C(n_{j-1}+n_j, n_j) = \frac{N!}{\prod_{j=1}^J n_j!}, \]
  where
  \[ n_1 + n_2 + \ldots + n_J = N. \quad (*) \]
- Each state $j$ has $g_j$ “degeneracy” levels (also independently occupied).
- Thus, the number of ways that there are $n_j$ different particles in energy state $j$, for all $j \in \{1, 2, \ldots, J\}$, is
  \[ W:= \prod_{j=1}^J g_j^{n_j} N! / \prod_{j=1}^J n_j! = N! \prod_{j=1}^J g_j^{n_j} / n_j!. \]
Some elementary applications to statistical mechanics – preliminaries (cont)

• Suppose that energy state \( j \) has associated energy value \( \varepsilon_j \).
• Also, that the total energy of the particles is fixed as \( Y > 0 \), i.e.,
  \[ n_1 \varepsilon_1 + n_2 \varepsilon_2 + \ldots + n_J \varepsilon_J = Y. \quad (***) \]
• Our objective is to find the most numerous (mode) arrangement of particles to energy states/degeneracy levels by minimizing
  \[ f_0(n) := -\log(W) = -\log(N!) - \sum_{j=1}^{J} n_j \log(g_j) + \sum_{j=1}^{J} \log(n_j!). \]
  (equivalently, maximizing \( A \)) subject to two equality constraints
  \[ f_1(n) := n_1 + n_2 + \ldots + n_J - N = 0 \]
  \[ f_2(n) := n_1 \varepsilon_1 + n_2 \varepsilon_2 + \ldots + n_J \varepsilon_J - Y = 0 \]
• So, define the Lagrangian \( L(n, \nu) \equiv f_0(n) + \nu_1 f_1(n) + \nu_2 f_2(n) \)

Maxwell-Boltzmann statistics

• By assuming \( N \) is large and also that for all \( j \), minimizing \( n_j \) is large, we can use Stirling’s approximation
  \[ \log(n!) = n \log(n) - n + o(n) \approx n \log(n) - n, \]
  which can be seen by upper Riemann sum approximation,
  \[ \log n! = \log(2) + \ldots + \log(n) \approx \int_1^n \log(x)dx = n \log(n) - n. \]
• After substituting Stirling’s approximation for each \( n_j \), the Lagrangian becomes
  \[ L(n, \nu) = -\log(N!) + \sum_{j=1}^{J} n_j (-\log(g_j) + \log(n_j) - 1 + \nu_1 + \nu_2 \varepsilon_j). \]
• Note that \( L \) is a convex function of \( n \geq 0 \).
• Relax the requirement that the \( n_j \) be integers, i.e., now \( n \in (\mathbb{R}^+)^J \).
Maxwell-Boltzmann statistics (cont)

- Solving the dual problem (i.e., ∀j, ∂L/∂n_j = 0) gives minimizing n as
  \[ n_j^* = g_j \exp(-v_1 - v_2 \varepsilon_j) \]
- So, \( G(\nu) = L(n^*, \nu) = -\log(N!) - \exp(-v_1) \sum_{j=1}^{J} g_j \exp(-v_2 \varepsilon_j) \).
- Choosing \( v_1 \) to meet \( n_1 + n_2 + \ldots + n_J = N \) \( (f_1 = 0) \), gives
  \[ n_j^* = g_j \exp(-v_1^* - v_2 \varepsilon_j) = N g_j \exp(-v_2 \varepsilon_j) / \sum_{m=1}^{J} g_m \exp(-v_2 \varepsilon_m) \],
  which again all need to be sufficiently large for Stirling’s approx.
- These are the Maxwell-Boltzmann statistics for \( \overline{n} \).
- For large N, substituting \( \overline{n}^* \) gives
  \[ \log(W) = v_1^* N + v_2 Y \]
- Boltzmann differentiated this expression, thereby deriving the second law of thermodynamics, and yielding
  \[ v_2 = 1/(kT) \],
  where T is absolute temperature and k is Boltzmann’s constant.

Bose-Einstein statistics for identical particles

- The number of ways putting \( n_j \) indistinguishable particles into \( g_j \) different degeneracy levels within an energy state \( j \), is \( C(n_j + g_j - 1, n_j) \).
- Here, the number of ways to assign particles, considering all energy states, is
  \[ W = \prod_{j=1}^{J} C(n_j + g_j - 1, n_j) \],
- Under particle number (*) and total energy (**) equality constraints, Stirling’s approximation gives as above:
  \[ n_j^* = g_j / (\exp(v_1 + v_2 \varepsilon_j) - 1) \],
  which are the Bose-Einstein statistics.
- Limiting one particle per energy state \( j \) leads to the Fermi-Dirac statistics, \( n_j^* = g_j / (\exp(-v_1 - v_2 \varepsilon_j) + 1) \).
Axiomatic probability

• In 1933, A.N. Kolmogoroff (Kolmogorov) used measure theory to generalize probability theory to accommodate stochastic processes of countably infinite duration (i.e., a discrete-time process made up of a countably infinite number of random variables), e.g., the stochastic process \( \{ T_1(n)/n, n \in \mathbb{N} \} \) on \( \Omega = \{1,2,\ldots,100\} \), where \( \mathbb{N} = \{1,2,3,\ldots\} \) is the set of natural numbers (strictly positive integers).

• Probability theory can also accommodate continuous-time stochastic processes (involving of abstract, uncountably infinite sample-spaces) thereby allowing for rigorous study of stochastic processes such as Brownian motion (and its derivative, white noise) in continuous time.

• We will explore Kolmogorov’s axiomatic probability theory and so rigorously define many of the terms we have mentioned previously.

• To that end, we first need to cover some basic ideas in set theory.

Review of set theory

• Suppose following sets are subsets of some “universal” set \( \Omega \).
• “x is an element of \( A \)” is denoted \( x \in A \)
• \( A \) is a subset of \( B \), denoted \( A \subseteq B \), when the following statement holds:
  \( \forall (\text{for all}) \ x \in \Omega, \text{ if } x \in A \text{ then } x \in B \)
• The intersection of two sets \( A \) and \( B \) is the set
  \[ AB = A \cap B = \{ s \in \Omega \ | \ s \in A \text{ and } s \in B \} = \{ s \in \Omega \ | \ s \in A, s \in B \} = BA \]
• The union of two sets \( A \) and \( B \) is the set
  \[ A \cup B = \{ s \in \Omega \ | \ s \in A \text{ or } s \in B \} = BU A \]
  where “or” is inclusive allowing \( s \in AB \), and the symbol “|” denotes “such that” (as does “:”).
• The complement of a set \( A^c = \{ s \in \Omega \ | \ s \notin A \} = \Omega - A \) (here need to specify \( \Omega \))
• Note that \( \Omega^c = \phi \), the empty set, and \( \phi^c = \Omega \).
• The difference of two sets is defined as
  \[ A - B = A \setminus B = \{ s \in \Omega \ | \ s \in A \text{ and } s \notin B \} = A \cap B^c = A - (AB) \]
• The exclusive union of two sets is defined as
  \[ A \oplus B = \{ s \in \Omega \ | \ s \in A \text{ or } s \in B \text{ but } s \notin AB \} = AUB - AB = (A-B) \cup (B-A) \]
• For example, if \( A = \{1,2\} \) and \( B = \{2,3,4\} \), then \( AUB = \{1,2,3,4\} \), \( AB = \{2\} \), \( A-B = \{1\} \), \( B-A = \{3,4\} \), \( A \oplus B = \{1,3,4\} \).
Venn diagrams for set operations/relations, where everything inside the box is $\Omega$

De Morgan’s theorem- a direct proof by exhaustive cases (truth table)

Theorem: $(A \cap B)^c = A^c \cup B^c$ \ \ $\forall$ sets A,B.

• Proof: for a particular $x \in \Omega$, define statements $p = \text{"x} \in A\text{"}$ and $q = \text{"x} \in B\text{"}$.
  • $p$, $q$ are either True=1 or False=0.
  • The logical negation operator is $\sim$, e.g., $\sim p = x \notin A$.
  • Note that $x \in AB \equiv p \land q$ (i.e., $p$ and $q$), $x \in A\cup B \equiv p \lor q$ (i.e., $p$ or $q$).
  • Proof of De Morgan’s theorem by truth table (all combinations of $p,q$):

<p>| | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$p$</td>
<td>$q$</td>
<td>$\sim p$</td>
<td>$\sim q$</td>
<td>$\sim(p \land q)$</td>
<td>$p \lor q$</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

• Note the logical equivalence (by column) $\sim(p \land q) \equiv \sim p \lor \sim q$.
  • Q.E.D.
De Morgan’s theorem - exercises

- **Exercise:** Show \((A \cup B)^c = A^c \cap B^c\) for all sets \(A, B\), as an immediate consequence (corollary) of the above half of De Morgan’s theorem by truth table.

- **Exercise:** Show the more basic “contrapositive”:
  \(A \subseteq B\) if and only if \(B^c \subseteq A^c\).

De Morgan’s theorem for equivalent AND (\(\wedge\)) and OR (\(\vee\)) gate configurations

\[
\begin{align*}
 p & \quad \rightarrow \\
 q & \quad \rightarrow \\
 \sim(p \land q) & \equiv \sim p \lor \sim q
\end{align*}
\]

\[
\begin{align*}
 p & \quad \rightarrow \\
 q & \quad \rightarrow \\
 \sim p \lor \sim q
\end{align*}
\]
Miscellaneous simple set identities

• Commutative property of $\cup$
  $A \cup B = B \cup A$ for all sets $A,B$

• Associative property of $\cup$
  $(A \cup B) \cup C = A \cup (B \cup C)$ for all sets $A,B,C$

• Similarly, associative and commutative properties of $\cap$

• Distributive property of $\cap$ over $\cup$:
  $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ for all sets $A,B,C$.

• Distributive property of $\cup$ over $\cap$:
  $(A \cap B) \cup (A \cap C) = A \cup (B \cap C)$ for all sets $A,B,C$.

• **Exercise**: prove these identities.

De Morgan’s theorem for arbitrary number of sets - Proof by induction

**Theorem**: $(A_1 \cup A_2 \cup \ldots \cup A_n)^c = A_1^c \cap A_2^c \cap \ldots \cap A_n^c$ for any sets $A_k$ and any $n \in \{2,3,4,\ldots\}$.

**Proof**:

• We have already established De Morgan’s rule for $n=2$.

• Inductive assumption: DeMorgan’s theorem for arbitrary (fixed) $n=k-1 \geq 2$ (or for all $n$ up to and including $k-1 \geq 2$ if needs be, i.e., strong inductive assumption).

• To complete the inductive proof, we need to establish De Morgan’s theorem for $n=k$.

• To this end, let $B = A_1 \cup A_2 \cup \ldots \cup A_{k-1}$

• Apply De Morgan with $n=2$: $(B \cup A_k)^c = B^c \cap A_k^c$ and substitute the inductive assumption on $B^c$.

• Q.E.D.
An Inclusion-Exclusion Formula

- An n-partition of a set B is a collection of n sets $\Delta_1, \Delta_2, \ldots, \Delta_n$ such that:
  - None are empty, $\Delta_k \neq \emptyset$ for all $k$
  - Covering, $B = \Delta_1 \cup \Delta_2 \cup \ldots \cup \Delta_n$
  - Disjoint (non-overlapping), $\Delta_k \cap \Delta_j \neq \emptyset$ for all $k \neq j$
- We can write an obvious inclusion-exclusion formula for the union of two arbitrary sets as the union of 3 disjoint sets (a disjoint covering):
  \[ A \cup B = [A-B] \cup [B-A] \cup [AB] \]
- Similarly, we can write the union of three arbitrary sets as disjoint-covering of 7 sets:
  \[ A \cup B \cup C = [A-(B \cup C)] \cup [B-(A \cup C)] \cup [C-(A \cup B)] \cup [AB-C] \cup [BC-A] \cup [CA-B] \cup [ABC] \]
- Ignoring the empty sets, these inclusion-exclusion decompositions (disjoint-coverings) form partitions of the left-hand-side unions of arbitrary sets.

Partitioning $\bigcup_{i=1}^{3} E_i$ into seven disjoint regions $(\Delta_1, \ldots, \Delta_7)$ by inclusion/exclusion:

**Exercise:** Write each set $\Delta$ in terms of the $E$ sets.
Inclusion-exclusion formula for arbitrary number of sets by induction

• Write $A_1 \cup A_2 \cup \ldots \cup A_n = B \cup A_n$ where $B = A_1 \cup \ldots \cup A_{n-1}$
• Find the 3-disjoint-covering of $B \cup A_n$ by inclusion-exclusion.
• Note that by De Morgan, the term to $A_n \setminus B = A_n \cap B^c$ simplifies to a single element of the inclusion-exclusion expansion.
• Apply the inductively assumed known inclusion-exclusion expansion for $B= A_1 \cup \ldots \cup A_{n-1}$ to each of the other two terms.
• Each of these other terms will produce $n-1$ elements after distribution of $A_n \cap (A_n^c)$ or $\cap A_n$
• Note that if $g(n-1)$ is the number of sets involved in the inclusion-exclusion formula for $n-1$ sets, then $g(n) = 2g(n-1) + 1$.
• Exercise: Fill in the details of this argument.
• Exercise: Does the argument work if you distribute $\cup A_n$ over the expansion of $B$ first, and then apply the 3-partition in each of the resulting terms?

An Inclusion-Exclusion Formula with Not-Disjoint Sets

• Consider four sets $A, B, C, D$ not necessarily disjoint.
• An inclusion-exclusion formula for
  \[ P(A \cup B \cup C \cup D) = P(A) + P(B) + P(C) + P(D) - P(AB) - P(AC) - P(AD) - P(BC) - P(BD) - P(CD) + P(ABC) + P(ABD) + P(ACD) + P(BCD) - P(ABCD) \]
• Note: $P(ABCD)$ is added $4 = \binom{4}{1}$ times in $P(A) + \ldots + P(D)$, $6 = \binom{4}{2}$ times in $\ldots - P(AB) - \ldots - P(CD)$, and added $4 = \binom{4}{3}$ times in $P(ABC) + \ldots + P(BCD)$, and thus must subtracted once so that it contributes only once to the union on LHS.
• Exercise: State and prove this inclusion-exclusion formula for arbitrary finite number of sets $n$ by induction.
An Inclusion-Exclusion Formula with Not-Disjoint Sets - example

- Find the probability that some player is dealt a complete suit in a game of bridge, with 4 players (A-D) evenly splitting the 52-card deck.
- \( P(A) \) := Probability that player A is dealt a complete suit - note that this event includes the possibility that B, C or D are also dealt a complete suit.
- So, by inclusion-exclusion, the answer is
  \[
  P(A \cup B \cup C \cup D) = P(A) + P(B) + P(C) + P(D) - P(AB) - P(AC) - P(AD) - P(BC) - P(BD) - P(CD) + P(ABC) + P(ABD) + P(ACD) + P(BCD) - P(ABCD)
  \]
- By symmetry, we get that the answer is
  \[
  C(4,1)P(A) - C(4,2)P(AB) + C(4,3)P(ABD) - C(4,4)P(ABCD)
  \]
  \[
  = 4P(A) - 6P(AB) + 4P(ABD) - P(ABCD)
  \]

Inclusion-exclusion – example (cont)

- \( P(A) = 4/C(52,13) \) because there are four suits (hence four 13-card flushes)
- \( P(AB) = 4\cdot3/[C(52,13)C(39,13)] \) because after A's flush is selected, B can select from 3 remaining flushes (numerator divided by 2! to account for overcounting).  
- \( P(ABC) = 4\cdot3\cdot2/[C(52,13)C(39,13)C(26,13)] \)
- \( P(ABCD) = 4![C(52,13)C(39,13)C(26,13)C(13,13)] \)
- Note \( P(ABCD) = P(ABC) \) because the last player is dealt the 13 remaining cards, i.e., one has either one, two or four players dealt a complete suit.
- After substituting, we get the final answer as
  \[
  4/C(52,13) - 72/[C(52,13)C(39,13)] + 72/[C(52,13)C(39,13)C(26,13)].
  \]
- Note that the denominator of the last term is the number of 4-player deals of the entire 52-card deck: \( 52!/(13!)^4 \) (after terms cancel out).
- Rozenov p. 24, Ch. 2, Prob. 16
Countable and Uncountable Sets

- A set is **countable** if each of its element can be placed in one-to-one correspondence with a subset of the positive integers, \( \mathbb{N} = \mathbb{Z}^{>0} \)
- A set is **countably infinite** if it is countable and has an infinite number of elements, e.g.,
  - the set of all integers, \( \mathbb{Z} \)
  - the set of all rational numbers, \( \mathbb{Q} \)
- **Uncountable** sets include:
  - the set of irrational numbers, \( \mathbb{R} - \mathbb{Q} \)
  - the set of real numbers, \( \mathbb{R} \)
- A very important property of the set of real numbers \( \mathbb{R} \) is its "separability", i.e., that it possesses a countable, dense subset \( \mathbb{Q} \).
- To see that \( \mathbb{Q}^+ \) is countable, create the matrix whose \((i,j)\)th entry is \( i/j \) and count the entries along anti-diagonals of increasing length.
- Exercise: modify this argument to show \( \mathbb{Q} \) is countable.

(c) 2013 George Kesidis

---

Probability Space

- Formally, a probability space \((\Omega, \mathcal{F}, P)\) is a triple consisting of
  - a sample space \( \Omega \) (of outcomes of a random experiment),
  - a \( \sigma \)-field (or \( \sigma \)-algebra) of events (subsets of \( \Omega \)) \( \mathcal{F} \), and
  - a probability measure \( P \).
- For a finite or countably infinite sample space, \( \mathcal{F} \) is often the set of all possible subsets of \( \Omega \) (including \( \emptyset \) and \( \Omega \)), i.e., the power set of \( \Omega \) denoted \( 2^\Omega \).
- The measure-theoretic aspects of a probability space, particularly the need to define a \( \sigma \)-fields of events for the case of uncountable sample spaces (rather than using the power set), are beyond the scope of this course.

(c) 2013 George Kesidis
Three axioms of a probability measure

1. \( P(A) \in \mathbb{R}^{\geq 0} \) for all events \( A \in \mathcal{F} \) (all probabilities are non-negative real numbers).
2. \( P(\Omega) = 1 \).
3. For any countable collection of disjoint events, \( A_1, A_2, A_3, \ldots \),
   \[ P(A_1 \cup A_2 \cup A_3 \ldots) = \sum_{n \geq 1} P(A_n), \]
   i.e., \( P \) is countably additive.

Three axioms of a probability measure – immediate consequences

- \( P(A^c) = 1 - P(A) \) for all \( A \in \mathcal{F} \)
- \( P(\emptyset) = 0 \)
- If \( A \subseteq B \) then \( P(A) \leq P(B) \)
- The probability that \( A \) or \( B \) occur,
  \[ P(A \cup B) = P(A-B) + P(B-A) + P(AB) \]
  \[ = P(A-B) + P(B) \]
  \[ = P(A) + P(B-A) \]
  and other similar additive identities from disjoint-coverings of unions by inclusion-exclusion.
- Exercise: Prove Boole’s inequality: for any countable set of events \( A_1, A_2, A_3, \ldots \),
  \[ P(A_1 \cup A_2 \cup A_3 \ldots) \leq \sum_{n \geq 1} P(A_n). \]
Independent events - definitions

• Events A and B are independent if the probability that A and B occur is the product of their probabilities, i.e., $P(AB) = P(A)P(B)$.

• A collection of events is said to be mutually independent (or just independent) if the probability of an intersection of any combination of them is the product of the probabilities of each event in that combination.

• A collection of events is said to be pairwise independent if all pairs of events from the collection are independent – this is a weaker notion than mutual independence.

Diagram of composition of balls in an urn
Independent events - example

• Recall again the previous figure representing the composition of balls in an urn.
• Balls are chosen uniformly at random from the urn in a random experiment. Let
  – A = event of choosing a black ball, P(A) = 10/30 = 1/3
  – B = event of choosing a light ball, P(B) = (5+10)/30 = 1/2
  – C = event of choosing an even ball, P(C) = (2+0+6+4)/30 = 2/5
• Note P(AB) = 5/30 = P(A)P(B) so A and B are indep.
• But P(AC) = 2/30 ≠ P(A)P(C) so
  – A and C are dependent
  – A, B and C are dependent, even though
    P(ABC) = 2/30 = P(A)P(B)P(C)

Conditioning on events

• If P(B) > 0, then the conditional probability of A given that B has occurred is denoted P(A|B).
• That is, P(A|B) has to do with the residual uncertainty that A has occurred given knowledge that B has occurred - given the outcome of the random experiment ∈ B what is the probability that the outcome is also ∈ A, i.e., is in ∈ AB.
  1. So, P(A|B) should increase with P(AB), i.e., the probability that both events A and B occur.
  2. Obviously, one expects that P(B|B) = 1, i.e., given that B occurred, the sample space of possible outcomes is “reduced” to B.
  3. If A and B are independent, it should be that P(A|B) = P(A), i.e., knowledge that B has occurred does not affect the uncertainty about whether A has occurred too.
• These three properties of conditional probability lead to the definition
  P(A|B) = P(AB)/P(B) when P(B) > 0.
Conditioning on events - example

- Consider again the game of poker with 5-card hands drawn from a deck of 52 different cards.
- The cards are sequentially drawn.
- Given the first two cards are aces, the probability that the hand will contain all four aces:
  \[ P(4 \text{ aces} | \text{first 2 cards are aces}) = \frac{\binom{2}{2} \binom{48}{1}}{\binom{50}{3}} = \frac{3 \cdot 2}{50 \cdot 49} \]
- Note how given the first two cards of the hand are aces, the “residual uncertainty” in the random experiment (of drawing a 5-card hand) is reduced to the 3 remaining cards.

Conditioning: disjoint events are not necessarily independent

- Given B, the sample space is transformed from \((\Omega, \mathcal{F}, P)\) to \((B, \{AB \mid A \in \mathcal{F}\}, P(\cdot \cap B|B))\).
- Again, if \(A\) and \(B\) are independent:
  \[ P(A) = P(AB)/P(B) = P(A|B) \]
- Note that if \(AB = \emptyset\) and \(P(A), P(B) > 0\) (i.e., disjoint, possible events), then \(A, B\) are not independent!
- Indeed, if \(AB = \emptyset\) and \(P(A), P(B) > 0\) and event \(A\) occurs, then event \(B\) cannot occur, i.e., \(P(B|A) = 0\), so \(A\) and \(B\) are dependent.
- Later, we will generalize this simple definition so as to be able to condition on the observation of a random variable.
Random variables

- A random variable is a measurable mapping $X: \Omega \rightarrow \mathbb{R}$.
- By “measurable” we mean that for all “interesting” subsets $B$ of $\mathbb{R}$ (i.e., $B$ a member of the Borel $\sigma$-field of $\mathbb{R}$), $X^{-1}(B) = \{x \in B \}$ is an event.
- The cumulative distribution function (CDF) of $X$ is defined as $F_X(s) = P(X \leq s) = P(X \in (-\infty, s])$.
- Note that $(-\infty, s] \subseteq \mathbb{R}$ (i.e., all real numbers $\leq s$) is a Borel set for all $s \in \mathbb{R}$.
- Indeed, the Borel $\sigma$-field is the smallest containing (i.e., is generated by) the sets $\{(-\infty, s] | s \in \mathbb{R}\}$.
- $P_X(B) := P(X \in B)$ is an probability measure (on the Borel subsets $B$ of $\mathbb{R}$) that is induced by the random variable $X$. 

$X^{-1}(B) = \{x \in B \} = E_B \subseteq \Omega$, $B = \{x(\zeta) | \zeta \in E_B \} \subseteq \mathbb{R}$, i.e., $X^{-1}(B)$ is an event.
Random variables - CDF

- Since $\pm \infty \not\in \mathbb{R}$ by and any random variable $X$ is only $\mathbb{R}$-valued by definition,
  - $F_X(-\infty) = P(\emptyset) = 0$
  - $F_X(\infty) = P(X \in \mathbb{R}) = P(\Omega) = 1$
- Also, since $r < s$ implies $\{X \leq r\} \subseteq \{X \leq s\}$, $F_X: \mathbb{R} \rightarrow [0,1]$ is nondecreasing.

Independent random variables

- Random variables $X, Y, Z$ are said to be (mutually) independent if the events $\{X \in A\}, \{Y \in B\}, \{Z \in C\}$ are independent for all Borel sets $A, B, C \subseteq \mathbb{R}$.
- A simpler test for independence follows.
- The joint CDF of $X, Y, Z$ is
  \[ F_{X,Y,Z}(r,s,t) = P(X \leq r, Y \leq s, Z \leq t), \quad r,s,t \in \mathbb{R}, \]
  recall that comma separated events are intersected.
- Since $\{Y \leq \infty\} = \Omega$, one can obtain a marginal CDF from the joint CDF as
  \[ F_X(r) = F_{X,Y}(r,\infty); \quad \text{also, } F_Y(s) = F_{X,Y}(\infty,s). \]
- By Dynkin’s $\pi/\lambda$-class theorem, $X$ and $Y$ are independent iff
  \[ F_{X,Y}(rs) = F_X(r) F_Y(s) \text{ for all } s,r \in \mathbb{R}. \]
- Note: There is also a weaker notion of pairwise independence of a group of random variables.
Example of throwing two dice

- Recalling the example of tossing two “fair” dice, let the random variables \( Y \) be the outcome of the first die and \( Z \) be the outcome of the second, i.e., \( Y(n,m) = n \) and \( Z(n,m) = m \), where \( n, m \in \{1, 2, 3, 4, 5, 6\} \).
- Note that \( \{Y = 1\} = \{(1,1),(1,2),..., (1,6)\} \), i.e., an event with six samples.
- Also, \( \{Y = 1, Z \in \{3,4\}\} = \{(1,3), (1,4)\} \) and \( \{Z \in \{3,4\}\} \) is an event with 12 samples.
- The two events \( \{Y = 1\} \) and \( \{Z \in \{3,4\}\} \) are independent because:
  \[
  2/36 = P(Y = 1, Z \in \{3,4\}) = P(Y = 1)P(Z \in \{3,4\}) = (6/36)(12/36)
  = (1/6)(2/6).
  \]
- Any two such events respectively involving \( Y \) and \( Z \) are independent, so \( Y \) and \( Z \) are independent random variables.
- That \( Y \) and \( Z \) are independent is a consequence of the assumption of the “fairness” of dice-tossing random experiment.

The law of total probability

- Suppose we have a countable event-partition \( \Delta_1, \Delta_2, ... \) of the sample space \( \Omega \).
- By additivity, \( \sum_{n \geq 1} P(\Delta_n) = 1 \).
- Moreover, let’s assume that \( P(\Delta_n) > 0 \) for all \( n \) (this is not a vacuous assumption as it is possible that non-empty events are impossible, i.e., have probability zero).
- Note that for any event \( B \), the events \( \{B \Delta_n \mid n \geq 1\} \) form a disjoint covering of \( B \).
- Thus, by additivity and the definition of conditional probability, we arrive at a law of total probability
  \[
  P(B) = \sum_{n \geq 1} P(B \Delta_n) = \sum_{n \geq 1} P(B \mid \Delta_n) P(\Delta_n).
  \]
Bayes’ theorem/rule

• Suppose we are given that an (observed) event $B$ has occurred (i.e., a “posterior” or “effect” event) and that we want to compute the probability that each of different possible “prior” (or “cause”) events $\Delta_k$ has also occurred, where the prior events partition the sample space; i.e., we want:

$$P(\Delta_k | B) = \frac{P(B | \Delta_k) \cdot P(\Delta_k)}{P(B)}$$

$$= \frac{P(B | \Delta_k) \cdot P(\Delta_k)}{\sum_{n \geq 1} P(B | \Delta_n) \cdot P(\Delta_n)}$$

• Bayes’ theorem (the second equality above), relates to the corresponding posterior probability $P(B | \Delta_k)$ which is typically easier to calculate through causal relationships.

• Also note how we have used the law of total probability to expand out the denominator.

Bayes theorem – example of a binary symmetric channel

• The following figure depicts a BSC with channel-input (source/transmitted) symbol $X \in \{0,1\}$ and channel-output (destination/received/decoded) symbol $Y \in \{0,1\}$ with 10% error rate.

• Suppose $P(X=1)=0.3$, so that $P(X=0)=0.7$, i.e., the transmitted symbol priors.

• Suppose at the channel output, the symbol 0 is received.

• A natural question is: What is the probability that the symbol 0 was transmitted (prior event $X=0$) given than the symbol 0 was received (posterior event $Y=0$).

• By Bayes’ theorem,

$$P(X=0|Y=0) = \frac{P(Y=0|X=0) \cdot P(X=0)}{[P(Y=0|X=0) \cdot P(X=0) + P(Y=0|X=1) \cdot P(X=1)]}$$

$$= \frac{0.9 \times 0.7}{[0.9 \times 0.7 + 0.1 \times 0.3]}$$

$$= \frac{21}{22} > 0.9, \text{ i.e., } P(X=0|Y=0)>P(X=0).$$

• Note how Bayes’ theorem allows us to directly use the causal information of the BSC (posterior probabilities) and the transmitted symbol priors (distribution of $X$), to compute $P(X=0|Y=0)$. 
Note: discrete random variables

- For an event $A \subseteq \Omega$, define the indicator Bernoulli{$0,1$} (2-valued) random variable
  $$1_A(\omega) = 1 \text{ if } \omega \in A, \text{ and } =0 \text{ else.}$$
- For a random variable $X$, define the events
  $$\Delta_k = \{ \omega \in \Omega \mid X(\omega) = k \} = X^{-1}(\Delta_k), \text{ i.e., } X=k \text{ on } \Delta_k.$$
- For the previous example of a random variable $X \sim \text{Bernoulli}{0,1}$,
  $$X(\omega) = 1\Delta_1(\omega), \omega \in \Omega.$$ 
- More generally, if $X \sim \text{Bernoulli}{a,b}$ for different scalars $a,b \in \mathbb{R}$, then
  $$X(\omega) = a1\Delta_a(\omega) + b1\Delta_b(\omega), \omega \in \Omega$$
  where $\Delta_b = \Omega - \Delta_a$, i.e., the complement of $\Delta_a$.
- Even more generally, for a “discretely” distributed random variable $X$ with countable strict range $R(X) \subseteq \mathbb{R}$,
  $$X(\omega) = \sum_{a \in R(X)} a1\Delta_a(\omega) \text{ with } P(X=a) = P(\Delta_a)>0 \text{ and}$$
  the set of events $\{\Delta_a \mid a \in R(X)\}$ forming a partition of the sample space $\Omega$
- We will return to random variables in the shortly.
The “base-rate fallacy”

- Suppose an intrusion/exploit $E$ may lead to an alert $A$.
- $P(E) > 0$ is the “base rate” of exploits.
- The “base-rate fallacy” is a situation where $P(E) > P(E|A)$, i.e., the unconditional probability of exploit is larger than the posterior probability of the exploit given the alert.
- Exercise: Prove that $1 > P(E) > P(E|A) > 0$ if and only if $P(A|E^c) > P(A|E)$.
- Conclude that the “base-rate fallacy” simply refers to a poorly designed alert that more likely indicates the absence of an exploit than the presence of one.
- A Neyman-Pearson objective is to design an alert that minimizes the probability of missed detection (false negatives) $P(A^c|E) = 1 - P(A|E)$ subject to an upper bound on false positives $P(A|E^c)$.

Probability tree diagrams

- There are many “conditioning” problems in probability that have plausible answers, arrived at by plausible arguments, that are both wrong, or sometimes only the argument is wrong.
- In such cases, clarification may be obtained through a diagram depicting the conditioning argument.
- In the following tree diagram,
  - the sum of the probabilities of daughter nodes (to the right) is equal to the probability of the attached parent node (to the left),
  - where the daughter nodes represent events forming a disjoint partition of the parent,
  - and the leftmost (root) node is the whole sample space, $\Omega$.
- So, all nodes at the same level (e.g., the leaf (rightmost) nodes) in the tree represent events forming a disjoint covering of $\Omega$.
- The edges connecting parent to child node/event may be labeled with conditional probabilities.
Probability tree diagram – example

• This “binary” tree diagram represents a two-step random experiment:
  – A 2-card hand is chosen at random from a standard 52-card deck.
  – Then a single card X is chosen from the hand.
  – We’re interested in the probability that X is an ace.
• Let $B_k$ be the event that $k$ aces are in the 2-card hand.

• The rightmost/leaf events are, e.g.,
  $B_1 \cap \{X=A\}$, i.e., they’re all different.
• $0.5 = P(X=A|B_1)$, $1 = P(X=A|B_2)$
• $P(B_0|\Omega) = P(B_0) = C(48,2)/C(52,2)$
• $P(X=A) = P(X=A|B_1)P(B_1) + P(X=A|B_2)P(B_2)$

Probability tree diagram – example (cont)

• Now suppose we’re given that $X$ is an ace and we want to find the probability that two aces are in the hand, i.e.,
  $P(B_2 | X=A)$
• By Bayes’ rule, we get
  $P(B_2 | X=A) = P(X=A|B_2)P(B_2)/P(X=A)$, where
  $P(X=A) = P(X=A|B_2)P(B_2) + P(X=A|B_1)P(B_1)$
  $+ P(X=A|B_0)P(B_0)$ (zero)
  $= P(B_2) + 0.5P(B_1)$
  $= \left(\frac{C(4,2)}{C(52,2)} + 0.5 \frac{C(4,1)C(48,1)}{C(52,2)}\right)$
• Exercise: Show that given that one of a pair of siblings is male, the probability that the other sibling is also male is 1/3 (not 1/2 !). Hint: Suppose 4 possible siblings, $\{f_1, f_2, m_1, m_2\}$, let $M$ be the number of males, and find $P(M=2|M=1)$. 

(c) 2013 George Kesidis
Two Envelopes Paradox

- Suppose a genie places $100 in one envelope and $200 in another, identical envelope and seals both so that their contents are hidden.
- The genie then distributes the envelopes uniformly at random between two people, A and B.
- A and B know that one envelope contains double the money of the other, but they do not know which envelope.
- Suppose person A has the option to switch envelopes with B.
- Let $X$ be the contents of the envelope given to A by the genie.
- After switching, person A can expect to receive
  \[ \frac{1}{2}(2X) + \frac{1}{2}(X/2) = 1.25X > X! \]
- That is, A can expect to profit from switching.
- What is wrong with this argument?
- Basically, the problem is incorrect conditioning.
- Many “articles” have been written about this “paradox”.

Let $Y$ be A’s envelope after switching.

\[
P(Y=200) = P(Y=200|X=200)P(X=200) + P(Y=200|X=100)P(X=100)
\]

\[
= 0 \cdot \frac{1}{2} + 1 \cdot \frac{1}{2} = \frac{1}{2}
\]

- Similarly, $P(Y=100) = \frac{1}{2}$
- So, $EY = 100 \cdot \frac{1}{2} + 200 \cdot \frac{1}{2} = 150 = EX$
The Monty Hall problem - background

• Monty Hall was the host of a TV game-show, Let's Make a Deal, where contestants in the audience dressed up in Halloween costumes and decided between prizes hidden behind doors or cash offered by Monty.
• Suppose only one valuable prize is hidden behind one of three doors and you, the contestant, want to select the door of the valuable prize.
• The Monty Hall (or Let's Make a Deal) problem in probability consists of three steps.
  1. You choose one of the doors (uniformly) at random.
  2. Monty then shows you one of the remaining two doors that has nothing (or something silly, like a goat or ostrich) behind it.
  3. You then decide whether to stick with your original choice or switch to the remaining unopened door.
• Question: Will you improve your chances of selecting the prize's door (i.e., winning) if you switch at step 3?
• The answer is yes, though the typical belief is that it will make no difference.
• This problem was the focus of an episode of "Myth Busters" because of the unexpected answer to this question.
• Apparently, Monty Hall himself never did this on the show: http://www.youtube.com/watch?v=c1BSkquWkDo

The Monty Hall Problem – Probability tree diagram with equally likely edges per parent

• A contestant is assumed either "always switch" or "never switch".
• Switching (black) leads to 3 loss (red) and 6 win (blue) scenarios, all equally likely.
• So, P(win | always switch) = 6/9 = 2/3 > P(win | never switch) = 3/9 = 1/3.
• Similarly, P(lose | always switch) = 3/9 = 1/3 < P(lose | never switch) = 2/3.
• Figure © Christopher Griffin.
Random variables – recall...

• A random variables is a measurable map \( X : \Omega \rightarrow \mathbb{R} \).
• By “measurable” we mean that for all “interesting” subsets \( B \) of \( \mathbb{R} \) (i.e., \( B \) a member of the Borel \( \sigma \)-field of \( \mathbb{R} \)),
  \( X^{-1}(B) = \{ X \in B \} = \{ \zeta \in \Omega \mid X(\zeta) \in B \} \) is an event (in \( \mathcal{F} \)).
• The cumulative distribution function (CDF) of \( X \) is
  \[ F_X(z) = P(X \leq z), \quad z \in \mathbb{R}, \]
  i.e., \( F_X : \mathbb{R} \rightarrow \mathbb{R} \) is nondecreasing, \( F_X(-\infty) = 0 \), and \( F_X(\infty) = 1 \).
• Note that the subset \((-\infty,z] \) of \( \mathbb{R} \) (i.e., all real numbers \( \leq z \)),
  is a Borel set for all \( z \in \mathbb{R} \) and \( \{ X \in (-\infty,z] \} = \{ X \leq z \} \).
• Indeed, the Borel \( \sigma \)-field is the smallest containing (i.e., is generated by) the sets \( \{ (-\infty,z] \mid z \in \mathbb{R} \} \).
Action of the random variable $X$

$X^{-1}(B) = \{ X \in B \} = E_B \subseteq \Omega, \quad B = \{ X(\zeta) \mid \zeta \in E_B \} \subseteq \mathbb{R}$

Expectation of a random variable

- A random variable $Y$ is said to be non-negative almost surely, i.e., $Y \geq 0$ a.s., if $P(Y \geq 0) = 1$.
- For our purposes, we can define the expectation (mean, average, or expected value) of $Y \geq 0$ a.s. as the Riemann-Stieltjes integral,
  \[ E_Y = \int_0^\infty s \, dF_Y(s) \approx \sum s \, P(Y=s) \text{ because} \]
  \[ P(Y \approx s) \approx P(s < Y \leq s + ds) = F_Y(s + ds) - F_Y(s) \approx dF_Y(s) \]
- We will give many examples of how to compute the Riemann-Stieltjes integral, e.g., the special case where $F_Y$ is differentiable leads to the ordinary Riemann integral
  \[ E_Y = \int_0^\infty s \, (F_Y)'(s) \, ds \]
- Note that an expectation can exist but be infinite.
- Also, for a (measurable) function $h: \mathbb{R} \rightarrow \mathbb{R}$, $h(Y)$ is a random variable with expectation
  \[ E[h(Y)] = \int_0^\infty z \, dF_{h(Y)}(z) = \int_0^\infty h(s) \, dF_Y(s) \]
- Exercise: Prove the last equality by a suitable change of variable of integration.
Expectation of a random variable (cont)

• If \( X \) is a signed random variable, we can decompose it into its positive and negative parts, i.e.,
\[
X = X^+ - X^-
\]
where
\[
X^+ = \max\{0, X\} \quad \text{and} \quad X^- = \max\{0, -X\}
\]
are a.s. non-negative random variables.

• The expectation of a signed random variable \( X \) is said to exist if
\[
E \max\{0, X\} < \infty \quad \text{or} \quad E \max\{0, -X\} < \infty,
\]
and if so
\[
E X = E \max\{0, X\} - E \max\{0, -X\}.
\]

• This definition avoids the "\( \infty - \infty \)" situation (cf. the Cauchy distribution).

• Again, if \( X \) is a random variable, then "measurable" functions of \( X \) are easily shown to be random variables too, e.g., \( \max\{0, X\} \) and \( \max\{0, -X\} \) are random variables.

Expectation of a random variable (cont)

• We can define joint CDFs of multiple random variables, e.g., for two random variables \( X \) and \( Y \),
\[
F_{X,Y}(x,y) = P(X \leq x, Y \leq y)
\]
where we remind that the comma here represents set (event) intersection or logical "and".

• For a measurable \( g: \mathbb{R}^2 \to \mathbb{R} \) and random variables \( X, Y \), we can define the expectation of the random variable \( g(X,Y) \) as
\[
E g(X,Y) = \iint g(x,y) \, dF_{X,Y}(x,y)
\]
where the double integration is over all of \( \mathbb{R}^2 \).
Expectation is a linear operator

- Expectation inherits the linearity property from integration.
- That is, for all constants/scalars \( a, b \in \mathbb{R} \) and random variables \( X,Y \) with finite expectation,
  \[ E(aX+bY) = aEX + bEY. \]
- **Exercise**: Prove this. Hint: We can easily obtain the marginal CDFs from the joint CDF, e.g.,
  \[ P(X \leq r) = F_X(r) = F_{X,Y}(r, \infty) = P(X \leq r, Y \leq \infty), \]
  where \( \Omega = \{ Y \leq \infty \} \) simply because \( Y \) is a random variable.

Moments of a random variable \( X \)

- The \( k \)th moment is \( E(X^k) \), the expectation of \( X^k \).
- Note that the \( k \)th moment is generally different from the first moment raised to the \( k \)th power, \( (EX)^k \).
- The \( k \)th centered moment is \( E(X - EX)^k \).
- The variance (about the mean) is the second centered moment,
  \[ \text{var}(X) = E(X - EX)^2 = E(X^2) - (EX)^2 \geq 0. \]
- **Exercise**: Verify the second equality. To do this, you will need to use the fact that: expectation is a linear operator, \( EX \) is a constant (scalar), and the expectation of a constant \( c \) is \( Ec = c \).
- Note that if we define \( Y = X - EX \), then by linearity
  \[ EY = EX - E(EX) = EX - EX = 0 \]
  i.e., \( Y \) is a centered” (zero mean) random variable.
- The variance is always non-negative, cf. Jensen’s or Cauchy-Schwarz inequalities.
- The standard deviation (about the mean) is \( \sigma(X) = \sqrt{\text{var}(X)} \geq 0 \), i.e., the square root of variance.
Moment generating function

- The moment generating function (MGF) is $m_X(z) = E \exp(zX)$ where $z$ is a scalar.
- The MGF captures the entire distribution of the random variable, whereas moments are usually only features.
- The MGF gets its name through the Taylor series expansion $\exp(zX) = 1 + zX + (z^2/2!)X^2 + (z^3/3!)X^3 + ...$, so that by linearity, $m_X(z) = 1 + zEX + (z^2/2!) E(X^2) + (z^3/3!) E(X^3) + ...$
- Thus, the distribution of a random variable is characterized by all of its moments.

Relating expectation to probability – Bernoulli random variables as event indicators

- An indicator of an event $A \subseteq \Omega$, $1_A$ or $1_A^\xi$:
  $1_A(\xi) = 1$ if $\xi \in A$ and $1_A(\xi) = 0$ if $\xi \notin A$, cf. Bernoulli random variables.
- $E1_A = 1 \cdot P(A) + 0 \cdot (1 - P(A)) = P(A)$
- If the event $A = \{X \in B\}$ for a random variable $X$ and Borel set $B \subseteq \mathbb{R}$, then we can compute $P(A) = P(X \in B) = \int_B dF_X(r) = \int_{\mathbb{R}} 1\{r \in B\} dF_X(r) = E1\{X \in B\}$.
- This extends to Borel sets in $\mathbb{R}^n$, e.g., for $B \subseteq \mathbb{R}^2$, $P( (X,Y) \in B ) = E 1\{(X,Y) \in B\} = \iint_B dF_{X,Y}(r,s)$
Discrete random variables

• The strict range of $X$ is the smallest set $R_X \subseteq \mathbb{R}$ such that $P(X \in R_X) = 1$.
• We may alternatively denote the strict range of $X$ as $R(X)$.
• For the constant random variable above, $R_X = \{c\}$.
• If $R_X$ is countable with only a finite number of elements in any finite interval of $\mathbb{R}$, then $X$ is said to be discretely distributed.
• Discretely distributed random variables have piecewise-constant CDF, $F_X(r) = P(X \leq r)$, $r \in \mathbb{R}$.

Constant/deterministic random variables

• Suppose $X$ is (almost surely) constant, i.e., there is a real number $c$ such that $P(X = c) = 1$.
• In this case, $F_X(s) = u(s-c)$, the Heaviside unit step shifted to $c$.
• Thus $dF_X(s) = \delta(s-c)ds$, where the Dirac delta-function (unit impulse): $\delta = u'$ (the derivative of the unit step).
• The impulse, an extremely important function for linear systems, has the following properties:
  • $\delta(s) = 0$ for all $s \neq 0$, and
  • $\int_{-\infty}^{\infty} \delta(s)g(s)ds = g(0)$ for all $g$ continuous at $0$ (sampling property).
• In particular, $\int_{B} \delta(s)ds = 1$ for any $B \subseteq \mathbb{R}$ that contains the origin in its interior (hence "unit" impulse).
• One can immediately show that $Eg(X) = g(c)$, $EX = c$ in particular.
• For constant random variables, $\text{var}(X) = 0$. 
Bernoulli random variables

• If $R_x$ has two elements, say $R_x = \{0,1\}$, then $X$ is said to be Bernoulli distributed.
• Let $q = P(X=0)$ and $1-q = P(X=1)$ with $0 < q < 1$.
• So the CDF $F_X(x) = q \ u(x) + (1-q) \ u(x-1)$, so that $dF_X(x) = q \ \delta(x) + (1-q) \ \delta(x-1)dx$.
• For example, fair coin is tossed and $X = 1\{\text{heads}\}$, i.e., $X=1$ if (i.e., indicates whether) the toss outcome is a head, otherwise $X=0$.
• So, $X$ is Bernoulli with $q=1/2$.
• Exercise: Show $E X = 1-q$, $\text{var}(X) = q(1-q)$,
  and $E e^{zX} = q + (1-q)e^z$
Probability Mass Function (PMF)

• The PMF of a discretely distributed (or just “discrete”) random variable $X$ is
  
  $$p_X(z) = P(X = z) > 0 \text{ for } z \in R_X (= R(X)).$$

• Obviously, $1 = \sum_{z \in R(X)} p_X(z)$.

• For $z \in R(X)$, $p_X(z) = F_X(z) - F_X(z^-) = P(X \leq z) - P(X < z) =$ size of the step/jump in $F_X$ at $z$.

• $F_X(s) = \sum_{z \in R(X)} u(s - z) p_X(z)$

• $F_X'(s) = \sum_{z \in R(X)} \delta(s - z) p_X(z) = \text{the probability density function (PDF) of a discrete random variable}$

Discrete random variables - Expectation

• For any (measurable) function $g$,
  
  $$E g(X) = \sum_{z \in R(X)} g(z) p_X(z)$$

• Since $R(X)$ is countable for discrete $X$, one can order the elements of $R(X) = \{x_i\}_{i=1}^{\infty}$ so that $x_i < x_{i+1}$ for all integers $i > 0$.

• So one can relate to the Riemann-Stieltjes integral defining expectation by writing $Eg(X)$ as:
  
  $$E g(X) = \sum_{i=1}^{\infty} g(x_i) p_X(x_i)$$
  
  $$= \sum_{i=1}^{\infty} g(x_i) (F_X(x_i) - F_X(x_{i-1}))$$
  
  $$= \int_{-\infty}^{\infty} g(z) dF_X(z)$$

where $x_0 := -\infty$ and recall $F_X(-\infty) = 0$. 
Example: (a) CDF of a discrete (non-uniform) RV $X$ with $EX=0\cdot0.2+1\cdot0.6+3\cdot0.2=1.2$; (b) PDF of $X$ (later).

Discrete random variables

- Note that if $R_X = \{0,1,2,\ldots\} = \mathbb{Z}_{\geq 0}$, then the moment generating function of $X$, 
  
  $$m_X(\theta) = \sum_{n=0}^{\infty} e^{\theta n} p_X(n) = E e^{\theta X},$$

  resembles the discrete-time Fourier transform of $p_X$.
- We will return to this observation when considering sums of independent random variables.
- For the jointly discrete case with $B \subseteq R_X \times R_Y$, we can use the joint PMF $p_{X,Y}(x,y) = P(X=x, Y=y)$ to compute 
  
  $$P( (X,Y) \in B) = E 1\{(X,Y) \in B\} = \sum \sum_{(x,y) \in B} p_{X,Y}(x,y)$$
Example discrete distributions

• We now give some examples of very well known discrete distributions involving just one or two parameters.
• We’ve already discussed the Bernoulli distribution.
• In the following, we will cover the following distributions:
  – Uniform
  – Binomial
  – Poisson
  – Geometric

Discrete Uniform Distribution

• If \( X \) is a discrete uniformly distributed random variable if its PMF \( p_X \) is constant over its finite-sized strict range, i.e., \( |R_X| < \infty \).
• Thus, \( p_X(z) = \frac{1}{|R_X|} \) for \( z \in R_X \) (otherwise \( p_X(z) = 0 \))
• So, \( \mathbb{E}X = \sum_{z \in R_X} z \cdot p_X(z) = \frac{\left( \sum_{z \in R_X} z \right)}{|R_X|} \), i.e., simply the average value of the elements of \( R_X \)
• For any set \( B \subseteq \mathbb{R} \), \( P(X \in B) \) is the ratio of the number of elements in \( B \cap R_X \) to the number of elements in \( R_X \), as above for our “counting” (permutations & combinations) examples:
  \[ P(X \in B) = \mathbb{E} \mathbf{1}\{X \in B\} = \sum_{z \in R_X} \mathbf{1}\{z \in B\} \cdot p_X(z) = \frac{|B \cap R_X|}{|R_X|} \]
Discrete Uniform Distribution - Examples

• Suppose \( X \sim \text{UNIF}\{3,4,5,6,7\} \), i.e., uniformly distr'd on \( R_X = \{3,4,5,6,7\} \).
• Thus, \( p_X(z) = 1/5 \) for \( z \in \{3,4,5,6,7\} \).
• \( \text{EX} = 3 \cdot (1/5) + 4 \cdot (1/5) + 5 \cdot (1/5) + 6 \cdot (1/5) + 7 \cdot (1/5) \)
  \[ = (3+4+5+6+7)/5 \]
  \[ = 5 \]
  i.e., the arithmetic mean of \( R_X \)
• \( \text{E}(X^2) = (3^2 + 4^2 + 5^2 + 6^2 + 7^2)/5 = 27 \)
• Thus \( \text{var}(X) = \text{E}(X^2) - (\text{EX})^2 = 27 - 5^2 = 2 \)
• Note that we can also compute the variance of \( X \) as
  \[ \text{var}(X) = \text{E}(X-\text{EX})^2 = \sum_{z \in R_X} (z-\text{EX})^2 p_X(z) \]
  \[ = (3-5)^2 + (4-5)^2 + (5-5)^2 + (6-5)^2 + (7-5)^2 )/5 = 2 \]
• Note that \( (X-\text{EX})^2 \) with strict range \( \{0,1,4\} \) is not uniform.
• Again, note that \( \text{var}(X) = 0 \) if and only if \( X \) is constant a.s.

Discrete Uniform Distribution – Examples (cont)

• Suppose \( Y \sim \text{UNIF}\{4,5,6\} \)
• Thus, \( p_Y(z) = 1/3 \) for \( z \in \{4,5,6\} \).
• Check that \( \text{EY} = 5 \), i.e., the same mean as \( X \).
• But \( \text{var}(Y) = 2/3 < \text{var}(X) \).
• That is, \( Y \) has less variation about its mean than \( X \), as expected since \( R_Y \subseteq R_X \)
• Finally, let’s compute
  \( P(-3<Y<4.5) = P(Y=4) = 1/3 \), where \( (-3,4.5) \cap R_Y = \{4\} \).
• Alternatively, define the interval \( B = (-3, 4.5) \):
  \[ P(-3<Y<4.5) = P(Y \in B) = E1\{Y \in B\} = \sum_{z \in R_Y} 1\{z \in B\} p_Y(z) \]
  \[ = 1\{3 \in B\} (1/3) + 1\{4 \in B\} (1/3) + 1\{5 \in B\} (1/3) \]
  \[ = 0 + 1/3 + 0 = 1/3 \]

(c) 2013 George Kesidis 117

(c) 2013 George Kesidis 118
The binomial distribution

- The binomial distribution $\text{BINOM}(n,q)$, with parameters $n$ and $q$ such that $n \in \{1,2,3,\ldots\}$ and $0 < q < 1$, has PMF:
  
  $$p(m) = C(n,m) \cdot q^m \cdot (1-q)^{n-m}$$

  for integers $m$ such that $0 \leq m \leq n$.

- The binomial coefficients form Pascal's triangle:
  
  \[
  \begin{array}{cccc}
  & C(0,0) & & \\
  C(1,0) & C(1,1) & & \\
  C(2,0) & C(2,1) & C(2,2) & \\
  C(3,0) & C(3,1) & C(3,2) & C(3,3) \\
  \end{array}
  \]

  where the leftmost and rightmost quantities are all 1, and every interior element is the sum of the element to the upper right and upper left, i.e., Pascal's identity: $C(n,m) = C(n-1,m) + C(n-1,m-1)$.

- **Exercise:** Use induction to prove Pascal's identity and the binomial theorem, $1 = \sum_{m=0}^{n} p_X(m)$.

- **Exercise:** Numerically evaluate Pascal's triangle for several rows.

---

CDF for a binomial RV with $n = 4$, $q = 0.6$
Binomial distribution from independent Bernoulli trials

- Suppose a coin is tossed and the outcome is tails with probability $q$, and so the outcome is heads with probability $1-q$.
- The coin is obviously unfair (biased) if $q \neq 0.5$.
- Consider $n$ independent tosses of such a coin, i.e., $n$ independent Bernoulli trials.
- The total number of tails can be written as $X = \sum_{k=1}^{n} B_k$, where $B_k := 1\{k^{th} \text{ toss is a tail}\}$ are i.i.d. Bernoulli random variables.
- We next determine the distribution of $X$.

Binomial distribution from independent Bernoulli trials (cont)

- For some integer $m$ such that $0 \leq m \leq n$, the number of elements/samples of the event $\{X=m\}$ is $C(n,m) = \text{the number of ways to obtain } m \text{ tails from } n \text{ coin tosses}$.
- The probability of each such sample is, by independence, $q^m(1-q)^{n-m}$.
- So, $X$ is binomially distributed with parameters $n,q$, denoted $X \sim \text{Binomial}(n,q)$, i.e.,
  $$p_X(m) := P(X=m) = C(n,m) q^m(1-q)^{n-m} \text{ for integers } 0 \leq m \leq n.$$
Mean and MGF of a Binomial RV

• Recall $X = \sum_{k=1}^{n} 1\{k^{th} \text{ toss is a tail}\} \sim \text{Binomial (n,q)}$

• We can use linearity (superposition specifically) to find
  
  $EX = \sum_{k=1}^{n} E 1\{k^{th} \text{ toss is a tail}\}$
  
  $= \sum_{k=1}^{n} P(k^{th} \text{ toss is a tail})$
  
  $= nq$

• Also, the MGF
  
  $m_X(z) = E \exp(zX) = E \exp(z\sum_{k=1}^{n} 1\{k^{th} \text{ toss is a tail}\})$
  
  $= E \prod_{k=1}^{n} \exp(z1\{k^{th} \text{ toss is a tail}\})$  ...indep. implies uncorrelated
  
  $= (qe^z + (1-q))^n$  ...by the MGF of identically distr’d Bernoulli’s

• Note how sums of independent random variables (in this case i.i.d. Bernoulli’s) are easily managed via MGFs.

The Binomial Theorem

• Using induction on $n$, it can be easily shown that
  
  $(a+b)^n = \sum_{m=0}^{n} C(n,m) a^m b^{n-m}$  ...recall binomial theorem exercise.

• As we argued for independent Bernoulli trials (sequence of coin tosses), note that
  
  $(a+b)^n = (a+b)(a+b)(a+b)...(a+b)$

• So, there are $C(n,m)$ different ways of choosing “a” $m$ times from these $n$ factors (and hence choosing “b” $n-m$ times), i.e., the coefficient of $a^m b^{n-m}$ is $C(n,m)$ in the expansion of $(a+b)^n$ above.

• Note that a Bernoulli distribution corresponds to $a=q\geq 0$ and $b=1-q\geq 0$ so that $a+b=1$.

• An alternative computation of the MGF of $X \sim \text{Binomial(n,q)}$ is:
  
  $E e^{Xz} = \sum_{m=0}^{n} e^{zm} p_X(m) = \sum_{m=0}^{n} C(n,m) (e^z q)^m(1-q)^{n-m}$
  
  $= (e^z q +1-q)^n$  ...by the binomial theorem.
Size of the Power Set

- The power set of a set A, denoted $2^A$, is the set of all subsets of A, including the empty set, $\phi$.
- For example, if $A=\{1,2\}$, then $2^A=\{\phi,\{1\},\{2\},A\}$

**Claim:** If $|A|<\infty$ (i.e., the size of the set A is finite, or A has finitely many elements), then the size of its power set is $2^{|A|}$, i.e., $|2^A|=2^{|A|}$.

**Proof:**
- The number of different subsets of size m is $C(|A|,m)$.
- So, $|2^A| = \sum_{m=0}^{[A]} C(|A|,m) = \sum_{m=0}^{[A]} C(|A|,m) 1^m 1^{[A]-m} = (1+1)^{|A|}$, where the last equality is by the binomial theorem.
- Q.E.D.

Example: Parity bit for error detection

- A parity bit may be appended to a data-packet of n bits in an attempt to detect bit-errors introduced in the transmission of the packet.
- Consider a parity bit $x$ generated by the transmitter by adding all n bits of the data packet and sets the parity bit to 1 if the sum is odd, otherwise the sum is 0.
- Upon receipt of $n+1$ bits (data packet plus parity), the receiver checks the sum of the (received) n bits against the parity; if they don't match, the receiver deems the packet received in error.
- Note that the single parity bit can detect only an odd number of bit-errors in transit.
Example: parity bit and binomial distribution

- Assume bit errors in transmission (including the parity bit!) occur independently and with probability $q$.
- The probability that at least one error in transit occurs is $1$ - the probability that no errors occur, i.e., $1 - (1-q)^{n+1}$.
- The probability that the receiver detects an error given that an error in transit occurred is (for $n$ even)
  \[
  P(\text{detect} \mid \geq 1 \text{ bit errors}) = \frac{[P(1 \text{ error}) + P(3 \text{ errors}) + \ldots + P(n+1 \text{ errors})]}{P(\geq 1 \text{ errors})}
  \]
  \[
  = \frac{[C(n+1,1)q(1-q)^n + C(n+1,3)q^3(1-q)^{n-2} + \ldots + q^{n+1}]}{[1 - (1-q)^{n+1}]}
  \]
- Note the use of the binomial distribution in the numerator.

Expectation – example using indicators

- Suppose that there are $n$ people with each with one of 365 different birthdays, where the birthdays are independent random variables (in particular, there are no twins in this population).
- **Exercise**: Show that the distribution of the number of birthdays $X_d$ on a particular day $d$ is Binomial with parameters $n$ and $q=1/365$, i.e., with mean $n/365$.
- The expected number of days $Y$ that are the birthdays of exactly three people (where $n \geq 3$) is
  \[
  EY = E \sum_{d=1}^{365} 1(X_d = 3) = \sum_{d=1}^{365} E1(X_d = 3)
  = \sum_{d=1}^{365} C(n,3)(1/365)^3 (364/365)^{n-3} = C(n,3)(364)^{n-3} (365)^{1-n}
  \]
- Note that the indicators in the sum are dependent random variables.
- **Exercise**: Verify that it’s difficult to find the distribution of $Y$.
- See Ross problem 7.21.
Multinomial distribution

- More generally, suppose that there are n independent trials $X_1, X_2, \ldots, X_n$ of the same random experiment having one of $k \geq 2$ outcomes ($k=2$ in the case of Bernoulli trials).
- Enumerating the outcomes of each trial 1, 2, ..., k, let $p_i = P(X=i)$ be the PMF of X and define the number of observed outcomes $i$, $M_i = \sum_{j=1}^n 1\{X_j = i\}$, for $i \in \{1, 2, \ldots, k\}$.
- Note that the $M_i$ are dependent since $n = M_1 + M_2 + \ldots + M_k$.
- For non-negative integers $m_1 + m_2 + \ldots + m_k = n$, to find the joint PMF $P(M_1 = m_1, M_2 = m_2, \ldots, M_k = m_k)$, consider the number of ways to choose $m_1$ trials from $n$ for outcome 1, then $m_2$ trials from $n-m_1$ for outcome 2, etc.: $C(n, m_1) \times C(n-m_1, m_2) \times C(n-m_1-m_2, m_3) \ldots \times C(m_k, m_k) = n! / (m_1! m_2! \ldots m_k!) \equiv C(n; m_1, m_2, \ldots, m_k)$.
- Note that the probability of $m_i$ outcomes $i$ is $p_i^{m_i}$.
- This leads to a multinomial distribution of random k-vectors of non-negative integers $(M_1, M_2, \ldots, M_k)$ whose components sum to $n$: $C(n; m_1, m_2, \ldots, m_k) \prod_{i=1}^k p_i^{m_i}$.

Poisson distribution

- A random variable $X$ is Poisson distributed with parameter $\lambda > 0$ if its strict range is $R_X = \{0, 1, 2, 3, \ldots\}$ and its PMF is $p_X(i) = \lambda^i e^{-\lambda} / i!$ for $i \in R_X$.
- Recall the Taylor series: $e^{x} = \sum_{i \geq 0} \lambda^i / i! \Rightarrow 1 = \sum_{i \geq 0} p_X(i)$.
- Check that the moment generating function (MGF) of $X$ is $m_X(z) = E(e^{zX}) = \sum_{i \geq 0} e^{zi} p_X(i) = \sum_{i \geq 0} (e^{z\lambda})^i / i! = \exp(\lambda e^{z} - \lambda)$.
- Similarly, check that the 1st and 2nd moments of $X$ are $E(X) = \lambda$ and $E(X^2) = \lambda^2 + \lambda$.
- So, $\text{var}(X) = \lambda = EX$.
- So, the mean is $\lambda$ and the standard deviation is $\lambda^{0.5}$.
- This distribution can be used to characterize the Poisson stochastic process, which is used to model shot noise, queueing systems, cash flow of insurance companies, etc.
- Exercise: Show that if $X \sim \text{Poisson}(\lambda)$, then $\forall n \in \mathbb{Z}^+, E(X^n) = \lambda E((X+1)^{n-1})$.
From binomial to Poisson: the law of small numbers

- Poisson's theorem (sometimes called the law of small numbers) states that the Binomial(n,q) distribution converges to the Poisson(λ) distribution as q→0 and n→∞ in such a way that nq→λ.
- To see this, recall that the MGF of a Binomial(n,q) distribution is 
  \[ m(z) = (qe^z + (1-q))^n. \]
- Substituting \( \lambda/q \) for \( n \) we get 
  \[ (qe^z + (1-q))^{\lambda/q} = (1+q(e^z - 1))^{\lambda/q} \]
- Finally because \( (1+x)^{1/x} \to e \) as \( x \downarrow 0 \) (here with \( x = q(e^z - 1) \)) we get 
  \( (1+q(e^z - 1))^{\lambda/q} \to \exp(\lambda(e^z - 1)) \) as \( q \downarrow 0 \) for all \( z \in \mathbb{R} \), i.e., the MGF of Poisson(λ).
- Thus, for small q and large n we can approximate the Binomial(n,q) distribution by the Poisson(λ) distribution with \( \lambda = nq \).

Poisson distr’n: Bayes’ rule example

- As a binomially distributed random variable can be constructed by counting IID Bernoulli RVs, so a Poisson RV can be constructed by counting IID exponential RVs (the exponential, closely related to the discrete geometric distribution discussed next, is also discussed below).
- Exercise: Read about the construction of the Poisson process.
- Consider a corpus of long documents each produced by a sole author.
- Suppose the number of typos in documents by author n is ~Poisson(λ_n).
- Also suppose that author n has written the fraction p_n of the documents where \( 1= \sum_n p_n \) and for this example, \( p_m = P(A=m) \) is the “prior” PMF.
- Given that a uniformly at random chosen document has t typos, the probability that the document was written by a particular author m is 
  \[ P(A=m|T=t) = \frac{P(T=t|A=m) \ p_m}{\sum_n P(T=t|A=n) \ p_n} \quad \text{by Bayes’ rule} \]
  \[ = \frac{(p_m \ \lambda^m e^{-\lambda} / m!)}{(\sum_n (p_n \ \lambda^n e^{-\lambda} / n!))} \]
A random variable $X$ is geometrically distributed with parameter $\lambda$, $0<\lambda<1$, if its strict range is $R_X = \{0,1,2,3,...\}$ and its PMF is $p_X(i) = (1-\lambda)\lambda^i$ for $i \in R_X$, where $\sum_{i \geq 0} \lambda^i = 1/(1-\lambda)$.

Check that the moment generating function (MGF) of $X$ is $m_X(z) = (1-\lambda)/(1-\lambda e^z)$ for $z$ such that $|\lambda e^z|<1$.

To compute $EX$, we rely on a little trick involving a derivative:

$$EX = \sum_{i \geq 0} i p_X(i) = (1-\lambda)\lambda \sum_{i \geq 1} i \lambda^{i-1} = (1-\lambda)\lambda \frac{d[\sum_{i \geq 0} \lambda^i]}{d\lambda} = (1-\lambda)\lambda \frac{d[1/(1-\lambda)]}{d\lambda} = \lambda/(1-\lambda)$$

Check that the CDF $F_X(i) = (1-\lambda^{i+1})u(i)$ for $i \in R_X$.

A geometric random variable is memoryless

**Theorem:** If $X \sim \text{geom}(\lambda)$, then for any $y \in \{0,1,2,3,...\}$ and $x \in \{1,2,3,...\}$,

$$P(X>y+x|X\geq y) = P(X>y+x)/P(X\geq y) = (1-F_X(y+x))/(1-F_X(y-1)) = \lambda^{y+1} / \lambda^y = \lambda^{x+1} = P(X>x),$$

where we note in the numerator that the event $\{X>y+x, X\geq y\} = \{X>y+x\}$ (first equality) and in the denominator that $\{X\geq y\} = \{X\leq y-1\}$ (second equality). QED.

• Suppose $X$ here models the lifetime of a lightbulb.

• Given that the lightbulb has been observed to have operated (lived) for $y$ units of time, the probability that it lives for an additional $x$ units is just $P(X>x)$ unconditionally, i.e., regarding its residual lifetime, the lightbulb has forgotten (has no memory) of its age.

• **Exercise:** Show that the geometric distribution is the only discrete distribution with this memoryless property.
Example: ALOHA random access local area networking

- Consider a time-slotted communication channel where at each unit of time a packet/frame of data may be transmitted.
- N independent time-synchronized stations share this channel and stations make independent packet-transmission decisions from slot-to-slot.
- Let q be the probability that a station attempts to transmit a packet in a time-slot.
- If two or more stations attempt to transmit in a time-slot, no packets are successfully transmitted due to interference.
- A packet is successfully transmitted if only one station attempts to transmit in a time-slot.
- What is the probability that a packet is successfully transmitted in a time?
- What choice of q maximizes throughput?
- Intuitively, if q is too small then slots are wasted because no station transmits; if q is too large then slots are wasted due to interference.

Example: ALOHA (cont)

- There are N samples in the event that a packet is successfully transmitted in a given slot, each corresponding to the packet being transmitted by one of the N different stations.
- The prob. of each such sample is prob. that one station transmits and the probability that all N-1 others don’t = product of these probabilities by assumption that the stations act independently = q(1-q)^{N-1}.
- So, the probability that a packet is successfully transmitted in a slot is Nq(1-q)^{N-1}, which is the average aggregate/channel throughput (in packets per time-slot) of the channel.
- Nq(1-q)^{N-1} (or the individual station throughputs q(1-q)^{N-1}) is maximized by choosing q=1/N.
- So, the maximum mean channel throughput is (1-1/N)^{N-1} \to 1/e as N\to\infty and the maximum mean throughput per station is 1/(Ne).
- Exercise: prove the last two statements.
- Exercise: show that max mean channel throughput for asynchronous, though still slotted, ALOHA is \frac{1}{2e}.

(c) 2013 George Kesidis
Example: ALOHA (cont)

• To find the distribution of the time (in slots) between successful packet transmissions in the channel, suppose that a packet is transmitted in slot 0 (i.e., “given” this).
• Let $X$ be the time (slot number) of the next transmission, i.e., the distribution of $X$ is the object of interest.
• Recall that we assume stations make independent packet transmission decisions from slot to slot ($X$ does not depend on the fact that a packet was transmitted at time zero).
• $X=1$ with probability $z \equiv Np(1-p)^{N-1}$.
• $X=2$ with probability $(1-z)z$ because with prob. $1-z$ a packet is not transmitted in slot 1 and with probability $z$ it is transmitted in slot 2.
• Similarly, $X=k \in \mathbb{Z}_{>0}$ with probability $(1-z)^{k-1}z$.
• So, $X-1$ is geometrically distributed with parameter $\lambda = 1-z$.
• **Exercise:** Find the probability of $k$ successful transmissions during a fixed interval of length $K$ time-slots, where $k \in \{0, 1, 2, \ldots, K\}$.

Minimum of two independent geometric random variables

**Theorem:** If $X \sim \text{geom}(\lambda)$ and $Y \sim \text{geom}(\mu)$ are independent random variables, then the random variable $\min\{X,Y\} \sim \text{geom}(\lambda \mu)$.

**Proof:**

• For all $z \in \{0, 1, 2, 3, \ldots\}$,

\[
P(\min\{X,Y\} \leq z) = P(\min\{X,Y\} > z) = 1 - P(X > z, Y > z) = 1 - P(X > z)P(Y > z) = 1 - \lambda^{z+1}\mu^{z+1} = 1 - (\lambda\mu)^{z+1},
\]

where the third equality is by assumed independence.

• So, $\min\{X,Y\}$ is geometric with parameter $\lambda\mu$.
• QED
Minimum of two independent geometric random variables - example

- Suppose there are two lightbulbs both turned on at time 0 and with geometrically distributed lifetimes, one (A) with mean lifetime 5 months and the other (B) with mean lifetime 10 months.
- Find the expected time when the first one of them burns out.
- First, the parameter \( \lambda \) of A satisfies \( \lambda/(1-\lambda)=5 \), i.e., \( \lambda=5/6 \).
- Similarly, the parameter \( \mu \) of B satisfies \( \mu/(1-\mu)=10 \), i.e., \( \mu = 10/11 \).
- So, the parameter of the minimum lifetime is \( \lambda\mu = 50/66 \).
- And so, the expected time that the first one of them burns out is \( \lambda\mu/(1-\lambda\mu)=3.125 \) months.

Probability that a particular random variable achieves the min. of independent geometrics

**Theorem:** If \( X \sim \text{geom}(\lambda) \) and \( Y \sim \text{geom}(\mu) \) are independent random variables, then
\[
P(X=\min\{X,Y\})=(1-\lambda)/(1-\lambda\mu).
\]

**Proof:**
\[
P(X=\min\{X,Y\})=P(X\leq Y)
= E(1[X\leq Y])
= \sum_{x=0}^{\infty} \sum_{y=x}^{\infty} 1[x\leq y] p_{X,Y}(x,y)
= \sum_{x=0}^{\infty} \sum_{y=x}^{\infty} p_X(x)p_Y(y) \quad \text{by independence}
= \sum_{x=0}^{\infty} (1-\lambda)\lambda^x \sum_{y=x}^{\infty} (1-\mu)\mu^y
= (1-\lambda) \sum_{x=0}^{\infty} \lambda^x \mu^x
= (1-\lambda)/(1-\lambda\mu)
\]
\*QED
Geometric random variables: Exercises

- **Exercise:** Similarly compute \( P(X > \min \{X, Y\}) \) and then verify that \( P(X = \min \{X, Y\}) + P(X > \min \{X, Y\}) = 1 \).
- **Exercise:** Use induction to generalize the previous two theorems to an arbitrary number of independent geometrics.
- **Exercise:** Apply the previous theorem to the previous numerical example.
- The characteristic Markov property of (discrete-time) Markov chains is based on these properties and the memoryless property of the geometric distribution.

---

**Table of Common Discrete Random Variables:**

**PMFs and CDFs**

<table>
<thead>
<tr>
<th>Family</th>
<th>PMF</th>
<th>CDF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bernoulli</td>
<td>( q^k(1 - p)^{n-k} u(k(n-k)} )</td>
<td>( ku(k) + p^n(k-1) )</td>
</tr>
<tr>
<td>Binomial</td>
<td>( \binom{n}{k} p^n q^n [u(k) - u(n-k)] )</td>
<td>( k &lt; 0, \sum_{k=0}^{n} \binom{n}{k} p^n q^n u(k) )</td>
</tr>
<tr>
<td>Poisson</td>
<td>( \frac{\mu^k}{k!} e^{-\mu} u(k) )</td>
<td>( \frac{\gamma(k+1, \mu)}{k!} u(k) )</td>
</tr>
<tr>
<td>Geometric</td>
<td>( pu^k q^k u(k) )</td>
<td>( p \left( \frac{1 - q^{k+1}}{1 - q} \right) u(k) )</td>
</tr>
</tbody>
</table>

\( u(k) = 1(k \geq 0) \) indicates \( k \in \mathbb{Z}^+ \), the non-negative integers (whole numbers)
\( u(k) - u(k-(n+1)) \) indicates \( k \in \{0, 1, ..., n\} \) for the binomial
Uniform PMF \( p(m) = 1/|\mathbb{E}| \) when \( |\mathbb{E}| < \infty \) and \( m \in \mathbb{E} \)
cf. law of small numbers relating binomial to Poisson (\( \mu^k \) not \( \mu k \))
Example discrete PMFs

Example: Birthday paradox

- Suppose a group of n (distinct) persons each has an independent birthday among the 365 possible ones.
- The probability that at least two share a birthday
  \[= 1 - \text{probability that none share a birthday}\]
  \[= 1 - \frac{(365!/(365-n)!)/365^n}{\text{permutations without replacement}/\text{permut. with repl.}}\]
- If \(n \geq 23\) then the probability that at least two share a birthday is greater than or equal to 0.5.
- The surprisingly small result (in n) is known as the birthday paradox.
- Exercise: The correct answer above is found by considering permutations. Explain why the correct answer is not
  \[1 - \frac{C(365,n)}{C(365+n-1,n-1)}\]
Example: Birthday attack

• Suppose a hacker wishes to have a domain name server (DNS) associate the domain name www.kesidis.com with one of his own 32-bit Internet Protocol (IP) addresses so that he can intercept some of the critically important correspondences that are directed to this site.

• The hacker simultaneously transmits \( q \) identical queries for www.kesidis.com to the targeted DNS server.

• Further suppose that the targeted DNS server naively forwards each query to an authoritative DNS server using i.i.d. transaction identifiers (used to authenticate the authoritative DNS server’s response) which are 16-bits long and not known to the hacker.

• Shortly thereafter, and before the authoritative DNS can reply, the hacker also transmits \( s \) responses to the targeted DNS spoofing those of the authoritative DNS.

• Each such response associates www.kesidis.com with the hacker’s chosen IP address and contains a guess at one of the transaction identifiers generated by the targeted DNS.

Example: Birthday attack (cont)

• **Exercise:** Assuming \( s = q \equiv n \), find the value of \( n \) so that the probability that a forwarded query and spoofed response have the same transaction identifier is 0.5, i.e., the probability that the hacker guesses correctly and thereby poisons the targeted DNS’s cache.
Exercise: “best of” games: Introduction

• Tennis is a game where the first man (resp. woman) to win three sets (resp. two sets) wins the match.

• We will simplify matters by not considering deuce (i.e., the need to win by at least two games or points) and tie-breakers by stipulating that:
  – a set is won by the first player to win 6 games, and
  – a game is won by the first player to win 4 (individual serve and volley) points, i.e., the first player to win the best of 7.

• Assume the outcomes of all points are mutually independent and identically distributed given the two players, i.e., don’t consider service advantages when reckoning the probability that a player will win a point.

Tennis: Dominant players only marginally better at points

• Suppose that player A is better than player B in that the probability q that A wins a point over B is such that q>0.5.

• Note that regarding service advantage,
  – we can let $p_X$ is the probability that player X wins a point on their service,
  – take $q = 0.5(p_A+1-p_B)$, and
  – assume players serve on the same number of points in a match.

• A tennis match is such that even if $q$ is very close to 0.5, e.g., $q=0.51$, the probability that A will win the match against B is very large.

• The very best players in the world are only marginally better than lower ranked players at the “point level” (q).
Tennis: Winning a game

- To see this, let’s compute the probability that A wins a game over B.
- A game is a sequence of independent and identically distributed (i.i.d.) Bernoulli trials with
  \( q := P(A \text{ wins}) \) and \( 1-q = P(A \text{ loses}, \text{i.e., B wins}) \)
- We want to find the probability that A wins 4 times before B does, i.e., that A wins the best of 7 i.i.d. points.
- The number of points \( N \) must therefore be between 4 and 7, i.e., \( 4 \leq N \leq 7 \), because
  - a minimum of four points are required for A to win, and
  - if A wins, then the most points that B can win is 3 (\( 4+3 = 7 \)).

Tennis: Winning a game (cont)

- By taking cases on \( N \), we get that
  \[
P(A \text{ wins game}) = \sum_{n \in \{4,5,6,7\}} P(A \text{ wins game, } N=n) = q \sum_{n \in \{4,5,6,7\}} C(n-1,3)q^3 (1-q)^{n-4} =: g
  \]
  where we have used the negative binomial distribution.
- To see the second equality, consider a sequence of \( n \) points that lead to a game won by A.
  - The last (\( n^{th} \)) point must be won by A (the initial factor \( q \)).
  - Of the remaining (\( n-1 \)) points, A must have 3 and B the rest (\( n-1-3=n-4 \) and \( C(n-1,3)=C(n-1,n-4) \)).
- Note the role of the binomial distribution on the point-sequence outcomes (coin-tosses) given \( N=n \).
Tennis: Winning a game (cont)

- We can easily generalize this to the case where \( k \) points must be won to win the game:
  \[
  W(q,k) := q \sum_{n \in \{kk+1, \ldots, 2k-1\}} C(n-1,k-1) q^{k-1} (1-q)^{n-k}
  \]
- Note that \( g = W(q,4) \).
- **Question (Hard):** Extend this result to a game where a player must win by two points (as in tennis with deuce for games and tie-breakers for sets).
  - That is, the first team to win \( n \) points wins the game (as in best of \( 2n-1 \)) except if both teams are tied at \( n-1 \) at which point the residual game is just the first team to go ahead by two points.
  - Hint: If the teams are tied at \( n-1 \), consider repeated trails (as in the discussion of geometric distributions) of successive pairs of independent points. The trials continue if the point-winner pattern is BA or AB, but eventually stops if the pattern is AA or BB. The question is: what is the probability that the terminating pattern will be AA instead of BB?

**Questions:** Prove

- \( \forall k \geq 1, W(0,k) = 0 \text{ and } W(1,k) = 1 \)
- \( \forall q, W(q,1) = q \)
- \( \forall k \geq 1, \text{ and } W(0.5,k) = 0.5 \)
- **Question (hard):** Prove the claim of the previous slide, that tennis games “boost” the likelihood of winning a point, i.e., \( \forall k > 1, W(q,k+1) > W(q,k) > q \) for all probabilities \( q > 0.5 \).
  **Hint:** Use induction on \( k \) with Pascal’s identity:
  \[
  \forall n > i > 0: C(n,i) - C(n-1,i-1) = C(n-1,i-1)
  \]
- **Question:** Also prove Pascal’s identity, which is evident from Pascal’s triangle of binomial coefficients.
Tennis: Game, Set and Match

- So, we can now readily compute,
  \[ P(A \text{ wins set}) = W(g,6) = W(W(q,4),6) =: s \]
  where, again,
  \[ q = P(A \text{ wins point}). \]
- And similarly
  \[ P(A \text{ wins men's match}) = W(s,3) \]
  \[ = W(W(g,6),3) \]
  \[ = W(W(W(q,4),6),3). \]
- Note that \( P(A \text{ wins women's match}) = W(s,2). \)
- **Question**: Numerically verify that \( P(A \text{ wins men's match}) \approx 0.61 \) and \( P(A \text{ wins women's match}) \approx 0.59 \) when \( q = 0.51 \). Note also that \( W(0.51,4) \approx 0.52 \) and that \( W(W(0.51,4),6) \approx 0.56 \).
- **Question (Hard)**: Prove \( \forall q > 0.5, W(q,k) \) is increasing in \( k \).

---

Baseball – Giants vs Tigers in 2012 WS

- After the Giants won the first two games of the 2012 World Series (WS), I heard on the radio that in 52 past WS where one team, A, won the first two games, that team A went on to win the (best of 7) WS 41 times.
- **Given only this information at this point in the WS** (never mind that the Giants actually went on to sweep the 2012 series), and again assuming that the future games have i.i.d. outcomes, let \( q \) be the probability that the Giants win a subsequent game of the WS.
- The probability that Giants will win 2 games before Tigers win 4 is
  \[ V(q) := q \sum_{n \in \{2,3,4,5\}} C(n-1,1) q^{n-2} (1-q) = 41/52 \approx 0.7885 \quad (*) \]
  where \( 41/52 \) is a statistic from historical data, and \( n \) is the game index after the given first two games.
- This likelihood \( V \) is slightly more general notion than that of \( W(q,k) \).
- Note that \( V(q) \) is increasing in \( q \) and \( V(0.5) = 13/16 = 0.8125 > 0.7885 \) so \( q < 1/2 \), but one may expect \( q > 1/2 \) given that the Giants (Team-A) won the first two games.
- **Question**: What is going on? Explain by delving into the game of baseball. Plot \( V(q) \) and numerically solve for \( q \) satisfying \( (*) \).
Example with recursion:
Choosing your noodles

• Consider a bowl with n noodle strands, each with 2 ends.
• Two noodle-ends are independently chosen uniformly at random from the bowl; if both ends belong to the same noodle, the noodle is removed from the bowl, otherwise the ends are joined to form a longer noodle which is returned to the bowl.
• This process is repeated until the bowl is empty (in n steps).
• Find the expected number $X_n$ of noodles $f(n) = E X_n$ drawn from the bowl.
• Clearly $1 \leq X_n \leq n$.

Example with recursion:
Choosing your noodles (cont)

• To obtain the answer, consider what happens in the first choice:
  – with probability $n/C(2n,2) = 1/(2n-1)$ the ends of the same noodle are chosen, i.e., $X_n = 1 + X_{n-1}$
  – with probability $1 - 1/(2n-1)$ the ends of the different noodles are chosen, i.e., $X_n = X_{n-1}$
• By the linearity property of expectation $E X_n = f(n)$,
  $f(n) = [1 + f(n-1)] 1/(2n-1) + f(n-1)(1 - 1/(2n-1))$
  $= 1/(2n-1) + f(n-1)$
• By recursive substitution,
  $f(n) = 1/(2n-1) + 1/(2n-3) + \ldots + 1/3 + 1$, using $f(1) = 1$.
• This question was asked during job interviews in 2011.
Continuous random variables

- A continuously distributed (or just “continuous”) random variable has a differentiable CDF,
  \[ f_X(x) = F'_X(x) \]
  where \( f_X \) is the probability density function (PDF) of \( X \).
- Clearly,
  \[ 1 = \int_{\mathbb{R}} f_X(x) \, dx = \int_{-\infty}^{\infty} f_X(x) \, dx = F_X(\infty) - F_X(-\infty) = 1 - 0, \]
  \[ P(X \in B) = \int_B f_X(x) \, dx, \text{ and} \]
  \[ F_X(z) = \int_{-\infty}^{z} f_X(x) \, dx = P(X \leq z) = P(X \in (-\infty, z]) = \mathbb{E}_1\{X \in (-\infty, z]\} \]
- Also, since \( F_X \) is non-decreasing, \( f_X = F'_X \geq 0. \)
- \( \mathbb{E}g(X) = \int_{\mathbb{R}} xf_{g(X)}(x) \, dx = \int_{\mathbb{R}} g(x)f_X(x) \, dx \)

Continuous random variables

- In the case of a continuous random variable \( X \), its strict \( R_X \) is the support of its PDF \( f_X \), i.e.,
  \[ R_X = \{ z \in \mathbb{R} \mid f_X(z) > 0 \}. \]
- For continuous random variables \( X \), since the CDF is continuous without any jump discontinuities, \( P(X = z) = 0 \) for all \( z \in \mathbb{R} \).
- However, since \( P(X \in B) = \mathbb{E}1\{X \in B\} = \int_B f_X(x) \, dx \), we see that it’s possible that \( P(X \in B) > 0 \) when \( B \cap R_X \neq \phi \) (i.e., \( B \) overlaps with \( R_X \)).
- Note: Using the qualifier “continuous” alone is a little confusing as continuity alone (of the CDF) does not suffice for differentiability.
Common Continuous Probability Densities and Distribution Functions

Table 2.4-2  Common Continuous Probability Densities and Distribution Functions

<table>
<thead>
<tr>
<th>Family</th>
<th>pdf $f_X(x)$</th>
<th>CDF $F_X(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Uniform $U(a,b)$</td>
<td>$\frac{1}{b-a} [u(x-a) - u(x-b)]$</td>
<td>$\begin{cases} 0, &amp; x &lt; a, \ \frac{x-a}{b-a}, &amp; a \leq x &lt; b, \ 1, &amp; b \leq x \end{cases}$</td>
</tr>
<tr>
<td>Exponential $\mu &gt; 0$</td>
<td>$\frac{1}{\mu} e^{-x/\mu} u(x)$</td>
<td>$\begin{cases} 0, &amp; x &lt; 0, \ 1 - e^{-x/\mu}, &amp; x \geq 0 \end{cases}$</td>
</tr>
<tr>
<td>Gaussian $\mathcal{N}(\mu, \sigma^2)$</td>
<td>$\frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2} \left( \frac{x-\mu}{\sigma} \right)^2 \right]$</td>
<td>$\frac{1}{2} + \text{erf}\left(\frac{x-\mu}{\sigma}\right)$</td>
</tr>
<tr>
<td>Laplacian $\sigma &gt; 0$</td>
<td>$\frac{1}{\sqrt{2\sigma}} \exp\left[-\sqrt{2}\frac{</td>
<td>x</td>
</tr>
<tr>
<td>Rayleigh $\sigma &gt; 0$</td>
<td>$\frac{x}{\sigma^2} e^{-x^2/2\sigma^2} u(x)$</td>
<td>$\left[1 - e^{-x^2/2\sigma^2}\right] u(x)$</td>
</tr>
</tbody>
</table>

Examples of Continuous Distributions
CDF of a continuous random variable $X$ uniformly distributed on $[0,T]$ 

- $f_X(t) = F_X'(t)$
- $F_X(t)$

$1/T$ if $T > 1$

$1$ if $T < 1$

Continuous Uniform Distribution

- If $X$ is a continuous uniformly distributed random variable if its PDF is constant on its strict range $R_X$ (i.e., over the support of the PDF).
- Thus, $f_X(z) = 1/|R_X|$ for $z \in R_X$ (otherwise $f_X(z) = 0$), where here $|R_X|$ represents the length (Lebesgue measure) of the set $R_X \subseteq \mathbb{R}$.
- So, $EX = \int_{\mathbb{R}} x f_X(x) dx = (\int_{R_X} x dx) / |R_X|$.
- For any set $B \subseteq \mathbb{R}$, $P(X \in B)$ is the ratio of the length of $B \cap R_X$ to that of $R_X$:
  
  $P(X \in B) = E 1\{X \in B\} = \int_{\mathbb{R}} 1\{x \in B\} f_X(x) dx = |B \cap R_X| / |R_X|$
Continuous Uniform Distribution - Example

- Suppose $X \sim \text{UNIF}[3,7]$, i.e., uniformly distr'd on the interval $R_X = [3,7]$.
- Thus, $f_X(z) = \frac{1}{(7-3)} = 1/4$ for $z \in [3,7]$.
- $EX = \int_{3}^{7} x \left( \frac{1}{4} \right) dx = \left( \frac{x^2}{8} \right)|_{3}^{7} = \frac{7^2}{8} - \frac{3^2}{8} = 5 = \text{midpoint of } [3,7]$.
- $EX^2 = \int_{3}^{7} x^2 \left( \frac{1}{4} \right) dx = \left( \frac{x^3}{12} \right)|_{3}^{7} = \frac{7^3}{12} - \frac{3^3}{12} = 79/3$
- Thus $\text{var}(X) = EX^2 - (EX)^2 = (79/3) - 5^2 = 4/3$.
- Note that we can also compute the variance of $X$ as $E(X-EX)^2 = \int_{3}^{7} (x-5)^2 \left( \frac{1}{4} \right) dx = \left( ((x-5)^3/12) \right)|_{3}^{7}$
  $= \frac{2^3}{12} - (-2)^3/12 = \frac{16}{12} = \frac{4}{3}$
- $P(-3 < X < 4.5) = P(X \in [3,4.5]) = \int_{3}^{4.5} (1/4) dx = (4.5-3)/(7-3) = 3/8$
- Finally, the CDF $F_X(z) = \int_{-\infty}^{z} f_X(x) \ dx$, so $F_X(z)=0$ if $z<3$, $F_X(z)=1$ if $z>7$, and $F_X(z) = (z-3)/4$ if $3 \leq z \leq 7$.

The Rayleigh, exponential, and uniform PDFs

![Graph of PDFs](image_url)

(c) 2013 George Kesidis
Exponentially distributed random variables

- $X \sim \text{exp}(\lambda)$ if $P(X > z) = e^{-\lambda z}$ for $z \geq 0$, where the parameter $\lambda > 0$.
- The exponential density is the continuous sister of the discrete geometric.
- $F_X(z) = (1 - e^{-\lambda z}) u(z)$ and $f_X(z) = F'_X(z) = \lambda e^{-\lambda z} u(z)$, where $u$ is the unit step, $u(z) := \mathbf{1}_{\{z \geq 0\}}$.
- $EX = 1/\lambda$ and $m_X(t) = \lambda/(\lambda - t)$ for $t < \lambda$.
- Exercise: Compute $\text{var}(X)$.

Exponentially distributed random variables are memoryless

- Exercise: show that exponentially distributed random variables are also memoryless, i.e.,
  $$P(X > y + x | X \geq y) = P(X > x) \text{ for all real } x, y \geq 0.$$ 
- Exercise: show that the exponential is the only continuously distributed random variable with this property.
- Exercise: show that if $X \sim \text{exp}(\lambda)$ and $Y \sim \text{exp}(\mu)$ are independent random variables, then
  - the random variable $\min\{X, Y\} \sim \text{exp}(\lambda + \mu)$, and
  - $P(X = \min\{X, Y\}) = \lambda / (\lambda + \mu)$.
- These properties can be used to construct continuous-time Markov chains.
Expectation for non-negative RVs
- integrating the complementary CDF

**Theorem:** If \( P(X \geq 0) = 1 \) (i.e., \( X \geq 0 \) almost surely (a.s.)), then
\[
EX = \int_0^\infty P(X > z) \, dz = \int_0^\infty (1-F_X(z)) \, dz
\]

**Proof:**
- \( EX = \int_0^\infty y \, dF_X(y) = \int_0^\infty \int_0^y dx \, dF_X(y) \)
- Switching the order of integration (Fubini’s theorem) gives
  \( EX = \int_0^\infty \int_x^\infty dF_X(y) \, dx = \int_0^\infty P(X > x) \, dx. \)
- Q.E.D.

**Example:** if \( X \) is exponentially distributed with parameter \( \lambda > 0 \), i.e., \( P(X > z) = \exp(-\lambda z) \) for all \( z \geq 0 \) and \( X \geq 0 \) a.s., then
\[
EX = \int_0^\infty \exp(-\lambda z) \, dz = 1/\lambda.
\]

**Exercise:** Repeat this calculation for the geometric distribution.

---

Use of Fubini’s theorem in the previous proof
Example: scheduling a server

• Suppose a CPU is waiting “on” since time 0 for a job that arrives at a random time $T > 0$ seconds with PDF $q$.
• Also suppose that while idling the CPU costs $r$ $$/s$, $b$ is the cost of turning the CPU, and $s$ is the time at which the CPU is scheduled to be turned off unless $T < s$ (the job arrives before).
• So if $s < T$, the CPU would need to be turned on to perform the job arriving at time $T$.
• The wasted expense is
  
  $$W(s) = rE \min \{T, s\} + bP(T < s)$$
  
  $$= r \int_0^s t q(t) dt + (rs + b) \int_s^\infty q(t) dt$$

Example: scheduling a server (cont)

• To find the $s^*$ that minimizes $W$, consider
  
  $$W'(s) = rsq(s) + r \int_s^\infty q(t) dt - (rs + b)q(s)$$
  
  $$= -bq(s) + r \int_s^\infty q(t) dt$$
• In the case where $T \sim \text{exp}(\lambda)$ where $ET = 1/\lambda$,
  
  $$W'(s) = -bq(s) + r \int_s^\infty q(t) dt$$
  
  $$= (-b\lambda + r)\exp(-\lambda s)$$
• So, if $r > b\lambda$, $W$ is increasing and take $s^* = 0$, else take $s^* = \infty$.
• If $T$ has a “Pareto-like” distribution with $P(T > t) = (t+1)^{-a}$, $a > 0$,
  
  $$W'(s) = (r - ba(s+1)^{-1})(s+1)^{-a}$$
• So, $W'(s^*) = 0$ when $s^* = \min \{0, -1 + ba/r\}$.
• Exercise: check that $W''(s^*) > 0$. 

(c) 2013 George Kesidis
Gaussian or normal distribution

- The Gaussian (or normal) distribution is extremely important primarily owing to central limit theorems, which will be covered shortly.
- The Gaussian distribution with strict range $\mathbb{R}$ is denoted $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is the mean and $\sigma \in \mathbb{R}^+ \setminus \{0\}$ is the standard deviation:
  
  \[
  f_X(x) = \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)/(2\pi\sigma^2)^{0.5}
  \]
  
  \[
  m_X(t) = \exp(t\mu+(t\sigma)^2/2)
  \]

- “Standard” normal is $N(0, 1)$, i.e., mean 0 and variance 1.
- Regarding the form of the MGF $m_X$, note that if $X \sim N(0,1)$ then $\mu+X\sigma \sim N(\mu, \sigma^2)$.

The Gaussian/Normal PDF
For Gaussian, the areas of the shaded region under curves are
(a) $P(X \leq x)$; (b) $P(X > -x)$; (c) $P(X > x)$; and for $x > 0$: (d) $P(-x < X \leq x) = P(|X| \leq x)$ and (e) $P(|X| > x)$.

From Poisson to Gaussian/normal –
a first central limit theorem

- Consider a random variable $Y_\lambda$ that is Poisson($\lambda$) distributed, i.e., $E Y_\lambda = \lambda$.
- Recall that $\text{var}(Y_\lambda) = \lambda$ so the standard deviation of $Y$ is $\lambda^{1/2}$.
- A central limit theorem for the Poisson distribution states that distribution of the random variable $X_\lambda \stackrel{def}{=} (Y_\lambda - \lambda) / \lambda^{1/2}$ converges to a standard normal.
- Re. standard normal, note that $E X_\lambda = 0$ and $\text{var}(X_\lambda) = 1$ for all $\lambda > 0$.
- To see this, recall the MGF of $Y_\lambda$ is $m_{Y_\lambda}(z) = \exp(\lambda e^z - \lambda)$.
- So, the MGF of $X_\lambda$ is $m_{X_\lambda}(z) = \exp(\lambda e^{z/\sqrt{\lambda}} - \lambda - z\lambda^{1/2})$
  $\Rightarrow \lim_{\lambda \to \infty} m_{X_\lambda}(z) = \exp(z^2/2)$, the Gaussian MGF.
- To see why, just substitute Taylor’s theorem applied to $e^{z/\sqrt{\lambda}} = 1 + z\lambda^{-1/2} + z^2\lambda^{-1} 2 + o(\lambda^{-1})$. 

(c) 2013 George Kesidis
173

(c) 2013 George Kesidis
174
From Poisson to Gaussian/normal – a first central limit theorem (cont)

• Convergence results of distributions to a Gaussian are known as central limit theorems (CLTs).
• Approximating distributions by Gaussians dramatically simplifies analysis and is a mechanism through which statistical confidence is assessed.
• So, we have shown how to go from Bernoulli (coin tossing) to binomial to Poisson (law of little numbers) to Gaussian (CLT) distributions.

The expectation of the Cauchy distribution does not exist

• PDF \( f(x) = \frac{1}{\pi(1+x^2)} \) for \( x \in \mathbb{R} \).
• Verify that \( f \) is a density.
• Verify that \( E(X1\{X \geq 0\}) = \infty \) and that \( E(-X1\{X \leq 0\}) = \infty \)
• Conclude that \( EX \) does not exist (since \( X \equiv X1\{X \geq 0\} + X1\{X \leq 0\} \)).
• Note that the question of existence is different from the question of finiteness.
• Exercise: Write python script to plot the density of the Cauchy. Note that it's also a "bell curve." Compare its "tails" to those of the standard Gaussian/normal, cf. Computer Exercise C.
Characteristic Function

- The characteristic function of a random variable $X$ is
  $\phi_X(v) = E e^{ivX}$ for $v \in \mathbb{R}$, where $i = (-1)^{0.5}$.
- So, for continuously distributed random variables, the characteristic function is the Fourier transform of the PDF:
  $\phi_X(v) = \int_{-\infty}^{\infty} e^{ivz} f_X(z) \, dz,$
  while the MGF is as the Laplace transform of the PDF.
- For a random variable $X$ is Laplace distributed, i.e.,
  $f_X(z) = 0.5ae^{-a|z|}, \ z \in \mathbb{R}$ and real parameter $a > 0$,
  $\phi_X(v) = 0.5a \int_{-\infty}^{\infty} e^{ivz} e^{-a|z|} \, dz$
  $= 0.5a \left[ \int_{-\infty}^{0} e^{(iv+a)z} \, dz + \int_{0}^{\infty} e^{(iv-a)z} \, dz \right]$
  $= 0.5a \left[ \frac{1}{iv+a} - \frac{1}{iv-a} \right]$
  $= \frac{a}{v^2+a^2}$

Characteristic Function – example (cont)

- Note how the characteristic function of the Laplace distribution is similar in form to that of the Cauchy distribution.
- Indeed, these two functions are Fourier transform pairs, for which there is a notion of duality through the inverse Fourier transform:
  $f_X(z) = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{ivz} \phi_X(v) \, dv$, $z \in \mathbb{R}$.
- For Laplacian $X$, this gives the identity for scalar $a > 0$:
  $0.5ae^{-a|z|} = (2\pi)^{-1} \int_{-\infty}^{\infty} e^{ivz} \left[ a/(v^2+a^2) \right] \, dv$, $z \in \mathbb{R}$.
- Take $a=1$ and swap the dummy variables $z$ and $v$ to get
  $e^{-|v|} = \int_{-\infty}^{\infty} e^{izv} \left[ 1/(\pi (z^2+1)) \right] \, dz$, $v \in \mathbb{R}$.
- Replacing $-v$ with $v$ gives the result that characteristic function of a Cauchy distributed random variable $Y$ is
  $\phi_Y(v) = \int_{-\infty}^{\infty} e^{ivz} \left[ 1/(\pi (z^2+1)) \right] \, dz = e^{-|v|}$, $v \in \mathbb{R}$.
- That is, the “forms” of the Laplace and Cauchy distributions are Fourier transform pairs.
- Exercise: Check by computing $\phi_Y$ by direct integration.
Densities for other types of distributions

- Random variables could have hybrid (mixed) discrete and continuous distributions.
- For the CDF $F$ depicted in the next figure, one can find a real number $0 < a < 1$ such that $F = a F_c + (1-a) F_d$, where $F_c$ is a continuous (i.e., differentiable) CDF and $F_d$ is a discrete (piecewise constant) CDF.
- There is a class of singular non-discrete (SND) distributions with uncountable strict-range (so not discrete) of (Lebesgue) measure zero (so non-differentiable/singular CDF).
- A SND CDF can be constructed using the Cantor set, an uncountable set of Lebesgue measure zero.
- Note: The expectation of SND random variables cannot be obtained via Riemann-Stieltjes integral, but rather can be defined more directly by Lebesgue integral on the sample space, $EY = \int_{\Omega} Y(s) \, dP(s)$

The CDF of a hybrid discrete-continuous random variable

![CDF of a hybrid discrete-continuous random variable](image)
A singular non-discrete distribution

- Take a random variable $U_0$ with uniform distribution on the interval $[0,1)$.
- For all Borel sets $B \subseteq [0,1]$, $P(U_0 \in B) = \lambda(B)$, where $\lambda$ is Lebesgue measure based on length of contiguous subintervals of $[0,1)$.
- Consider the ternary (base 3) representation of real numbers in $[0,1)$, i.e., using digits $\{0,1,2\}$.
- We will construct a sequence of uniform random variables $U_k$ for $k > 0$.
- $U_1$ is uniform on the interval $C_1 := [0, 1/3) \cup [2/3, 1)$ which is $[0,1)$ (support of $U_0$) with the middle third removed, i.e., elements of $C_1$ do not have digit 1 in the first position of their ternary representation.
- $U_2$ is uniform on the interval $C_2 := [0, 1/9) \cup [2/9, 1/3) \cup [2/3, 5/9) \cup [8/9, 1)$ which is $C_1$ (support of $U_1$) with the middle thirds of its intervals removed, i.e., elements of $C_2$ do not have digit 1 in either the first or second position of their ternary representation.
The Cantor-set distribution

- The set \( C = \lim_{k \to \infty} C_k \) is the Cantor set.
- The (uniform) CDFs \( F_k \) of \( U_k \) converge pointwise to CDF \( F \) whose support is \( C \) – this is convergence of the random variables \( U_k \) in distribution.
- The Lebesgue measure of \( C \) is
  \[
  1 - 1/3 - 2/9 - 4/27 - 8/81 \ldots = 1 - (1/3)(1 + 2/3 + 4/9 + 8/27 + \ldots) \\
  = 1 - (1/3)/(1-2/3) \\
  = 0
  \]
- Thus, \( C \) has zero “length” (i.e., \( F \) is a singular distribution).
- But the set \( C \) consists of all numbers in \([0,1)\) whose ternary expansion consists only of digits \( \{0,2\} \).
- Thus, \( C \) is uncountably infinite in size (i.e., \( F \) is not a discrete distribution) as there is a one-to-one mapping between its elements and the binary base-2 representation of elements whole interval \([0,1)\) !
- **Exercise:** Plot the CDFs \( F_k \) together, for \( k = 1,2,3,4 \). Can their limit \( F \) be drawn?

Recall that \( E_B = X^{-1}(B) \) in this example:
Functions of a random variable

- Recall for any (deterministic) measurable function $g$ of a random variable $Y$,
  \[ E_g(Y) = \int_{-\infty}^{\infty} g(s) \, dF_Y(s) \]

- To find the distribution of $g(Y)$ in terms of $F_Y$, define the inverse set relation
  \[ g^{-1}(B) = \{ y \in \mathbb{R} \mid g(y) \in B \} \text{ for all Borel } B \subseteq \mathbb{R} \]

- So, $F_{g(Y)}(x) = P(g(Y) \leq x) = P(Y \in g^{-1}(-\infty, x]) = \int_{g^{-1}(-\infty, x]} dF_Y(s)$

- If $g$ is strictly increasing, then
  \[ g^{-1}(\{x\}) \text{ is singleton set for all elements } x \text{ in the strict range } R_g \text{ of } g, \]
  \[ g^{-1} \text{ is also a strictly increasing function such that } g(g^{-1}(x)) \equiv x \text{ for all } x \in R_g, \text{i.e., } g^{-1} \text{ is just } g \text{ reflected in the line } y=x. \]

Functions of a random variable - example

- Suppose $Y \sim \text{exp}(3)$ and $X = 5e^{-7Y}$ and we need to find the PDF of $X$ and $EX$.

- $F_X(x) := P(X \leq x) = P(5e^{-7Y} \leq x) = P(Y \geq -(1/7)\log(x/5))$
  \[ = \exp((3/7)\log(x/5)) = (x/5)^{3/7} \text{ if } 0 < x \leq 5, \]

- $F_X(x) = 0$ for $x \leq 0$, and $F_X(x) = 1$ for $x > 5$.

- Thus, $f_X(x) = F_X'(x) = (3/35)(x/5)^{-4/7}$ if $0 < x \leq 5$, and $f_X(x) = 0$ else.

- $EX = \int_{-\infty}^{\infty} x f_X(x) \, dx = \int_0^5 x (3/35)(x/5)^{-4/7} \, dx$
  \[ = (3/7)5^{-3/7} \int_0^5 x^{3/7} \, dx = (3/10)5^{-3/7} x^{10/7} \bigg|_0^5 \]
  \[ = (3/10)5^{-3/7} (5^{10/7} - 0) = 3/2 \]

- Note: $X=g(Y)$ where decreasing $g(y)=5e^{7y}$
  and check that $EX = \int_{-\infty}^{\infty} g(y) f_Y(y) \, dy$

- So, $Y=g^{-1}(X)$ where $x = 5e^{-(7g^{-1}(x))}$
  \[ \Rightarrow g^{-1}(x) = -(1/7)\log(x/5) \]
Sampling an arbitrary distribution from a uniform

- If \( g \) is non-decreasing and right continuous, \( g^{-1} \) is also a non-decreasing and right-continuous function from \( \mathbb{R} \) to \( \mathbb{R} \), where \( g^{-1} \) is \( g \) reflected in the line \( y = x \) so that \( g^{-1}(g(x)) = x \equiv g(g^{-1}(x)) \).
- Suppose we want to generate a sample from an arbitrary CDF \( F_X \) (of a random variable \( X \)).
- **Theorem:** \( X \sim F_X^{-1}(U) \) where \( U \) is uniformly distributed over the interval \([0,1]\).
- **Proof:**
  - Note that \( F_X^{-1} \) is monotonically non-decreasing since any CDF \( F_X \) is monotonically non-decreasing.
  - Also, the domain of \( F_X^{-1} \) is \([0,1]\) = the range of \( F_X \).
  - So, \( P(F_X^{-1}(U) \leq z) = P(U \leq F_X(z)) = F_X(z) \).
  - Q.E.D.

For example, if \( X \sim \text{exp}(\lambda) \) then its CDF is \( F(x) = 1 - \exp(-\lambda x) \) for \( x \geq 0 \).
- Thus, \( y = 1 - \exp(-\lambda F^{-1}(y)) \) for \( 1 \geq y \geq 0 \), which implies \( F^{-1}(y) = -(1/\lambda)\log(1-y) \) for \( 1 \geq y \geq 0 \).
- So, \( X \sim -(1/\lambda)\log(1-U) \sim -(1/\lambda)\log(U) \).
- The figure below is for \( \lambda = 2 \).
Sampling an arbitrary distribution from a uniform – discrete example

• Now suppose X is discretely distributed on the state space \{a_1, a_2, a_3, a_4 \ldots\} with \(a_k < a_{k+1}\)
• The PMF of X is given by \(p_k = \Pr(X = a_k)\)

Functions of two or more random variables: joint distributions

• Recall that for \(g: \mathbb{R}^2 \to \mathbb{R}\) and random variables \(X, Y\), we can define the expectation of the random variable \(g(X,Y)\) as
  \[
  \mathbb{E} g(X,Y) = \iint g(x,y) \, dF_{X,Y}(x,y)
  \]
  where the integration is over all of \(\mathbb{R}^2\).
• For \(B \subseteq \mathbb{R}^2\), we can use this expression to compute \(\Pr((X,Y) \in B)\) by taking \(g(x,y) = 1_{\{(x,y) \in B\}}\), i.e., restricting the integration to over B.
• Also recall how we can obtain the marginal CDFs from the joint CDF, e.g.,
  \[
  \Pr(X \leq x, Y \leq y) = F_{X,Y}(x,y) = F_{X,Y,Z}(x, y, \infty) = \Pr(X \leq x, Y \leq y, Z \leq \infty),
  \]
  where \(\Omega = \{Z \leq \infty\}\) simply because \(Z\) is a random variable.
Buffon’s needle

• The following example involves a joint distribution of two *independent*, uniformly distributed random variables.

• Consider a plane ruled with east-west lines equally spaced at $A$ metres apart onto which a needle (line segment) of length $a$ metres is dropped “at random”.

• That is, the distance $X$ between the southern most tip of the needle and the nearest southern east-west line, and the acute angle $\Theta$ between the needle and south-north are
  – independent and
  – uniformly distributed, resp. $X \sim \text{UNIF}[0,A]$ and $\Theta \sim \text{UNIF}[-\pi/2,\pi/2]$.

• We’re interested in computing the probability that the needle will cross an east-west line,

$$P(X + a \cos(\Theta) > A)$$

Buffon’s needle (cont)

• Since $X \sim \text{UNIF}[0,A]$ is independent of $\Theta \sim \text{UNIF}[-\pi/2,\pi/2]$,

$$P(X + a \cos(\Theta) > A) = \int_{\max[0,A-a]}^{A} \int_{\arccos((A-x)/a)}^{\arccos((A-x)/a)} f_{\Theta,X}(\theta,x) \, d\theta \, dx$$

where $\forall \, z \in [0,1]$ we define $\arccos(z)=\cos^{-1}(z) \in [0,\pi/2]$, and

$$f_{\Theta,X}(\theta,x) = f_\Theta(\theta)f_X(x) = \frac{1}{\pi A} \text{ when } x \in [0,A] \text{ and } \theta \in [-\pi/2,\pi/2].$$

• Thus,

$$P(X + a \cos(\Theta) > A) = \int_{\max[0,A-a]}^{A} 2\cos^{-1}((A-x)/a) \, dx \, /\, (\pi A)$$

$$= \int_{\min[A/a,1]}^{\min[A/a,1]} 2\cos^{-1}(z) \, dz \, a/\, (\pi A)$$

$$= 2\cos^{-1}(z) - (1-z^2)^{0.5} \int_{\min[A/a,1]}^{\min[A/a,1]} 2a/\, (\pi A)$$

• So, when $a \leq A$, $P(X + a \cos(\Theta) > A) = 2a/(\pi A)$

• **Exercise:** Compute for the case of a long needle, $a > A$.

• **Exercise:** Explain how an estimate of $\pi$ can be obtained through repeated independent and identically distributed trials of dropping a short needle onto a hardwood floor.
Working with joint distributions

- Now consider the joint distribution of possibly dependent random variables.
- Preliminary operations with joint distributions involve techniques of
  - marginalization and
  - Integration (jointly continuous random variables) or summation (jointly discrete random variables) in multiple dimensions.
- In the jointly continuous case, the joint PDF of X, Y is
  \[ f_{X,Y}(x,y) = \frac{\partial^2 F_{X,Y}(x,y)}{\partial x \partial y} \]
- So, \( E g(X,Y) = \iint g(x,y) f_{X,Y}(x,y) \, dx \, dy \)
- In the jointly discrete case, the joint PMF of X, Y is
  \[ p_{X,Y}(x,y) = P(X=x, Y=y) \]

Point set associated with event \( A = \{ X \leq x, Y \leq y \} \), i.e., can evaluate \( P(A) = F_{X,Y}(x,y) = \int_{-\infty}^{x} \int_{-\infty}^{y} dF_{X,Y}(x',y') \)
\[ = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(x',y') \, dy' \, dx' \] for continuously distr’d r.v.’s
Point set for $A=\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$, where

$$P(A)=\int_{x_1}^{x_2} \int_{y_1}^{y_2} f_{X,Y}(x',y') dy' dx' = F_{X,Y}(x_2,y_2) - F_{X,Y}(x_2,y_1) - F_{X,Y}(x_1,y_2) + F_{X,Y}(x_1,y_1);$$

ex.: prove this

Points sets of events $A$ and $B$ –

$P(A)$ requires taking 3 cases in $\mathbb{R}^2$ to compute by integration
for $Y \geq 0$ a.s., $A = \{|X| \leq Y \leq 1\}$ so

$$P(A) = \int_0^1 \int_{-x}^x f_{X,Y}(x',y') \, dy' \, dx'$$

$$= \int_0^1 \int_y^1 f_{X,Y}(x',y') \, dx' \, dy' + \int_{-1}^0 \int_{-y}^1 f_{X,Y}(x',y') \, dx' \, dy'$$

**Figure 2.6-9** Shaded region in (a) to (e) is the intersection of $\text{supp}(f_{XY})$ with the point set associated with the event $\{-\infty < X \leq x, -\infty < Y \leq y\}$. In (f), the shaded region is the intersection of $\text{supp}(f_{XY})$ with $\{X + Y \leq 1\}$. 
The jointly Gaussian distribution

• Assume $X_i$ is Gaussian (normally) distributed with mean $\mu_i$ and variance $\sigma_i^2$, i.e., $X_i \sim N(\mu_i, \sigma_i^2)$.

• If independent RVs, the MGF of $X_1 + X_2$ is
  
  $m(\theta) = \exp(\mu_1 \theta + \sigma_1^2 \theta^2/2)\exp(\mu_2 \theta + \sigma_2^2 \theta^2/2)$ ...by independence
  
  which we also recognize as a Gaussian MGF.

• Even if dependent, $\alpha_1 X_1 + \alpha_2 X_2$, for (real) scalars $\alpha_i$, is Gaussian distributed with mean $\alpha_1 \mu_1 + \alpha_2 \mu_2$ and variance

  $\alpha_1^2 \sigma_1^2 + \alpha_2^2 \sigma_2^2 + 2\alpha_1 \alpha_2 \text{cov}(X_1, X_2),$

  where the covariance

  $\text{cov}(X_1, X_2) := E(X_1 X_2) - E(X_1) E(X_2) = E((X_1 - E(X_1))(X_2 - E(X_2)))$

  and recall $E(X_1 X_2)$ is the correlation of $X_1$ and $X_2$.

The jointly Gaussian distribution (cont)

• $X = (X_1, X_2, ..., X_n)^T$ are jointly Gaussian if their joint density

  $f_X(x) = \exp[-(x-EX)^T C^{-1} (x-EX)/2] / [(2\pi)^{n/2} \det(C)^{1/2}]$, $x \in \mathbb{R}^n$,

  where the (symmetric) covariance matrix is $C = E(X - EX)(X - EX)^T$ is assumed to be nonsingular here.

• For the case $n = 2$:

  $EX = [EX_1, EX_2]^T$ so that $(x-EX)^T = [x_1 - EX_1, x_2 - EX_2]^T$

  and the symmetric $2 \times 2$ covariance matrix

  $$C = \begin{bmatrix}
  \sigma_1^2 = \text{var}(X_1) & \text{cov}(X_1, X_2) \\
  \text{cov}(X_1, X_2) & \sigma_2^2 = \text{var}(X_2)
\end{bmatrix}$$

  so that

  $f_X(x_1, x_2) = \exp\left[-(x_1 - \mu_1)^2/2\sigma_1^2 + \rho(x_1 - \mu_1)(x_2 - \mu_2)/(\sigma_1 \sigma_2) + (x_2 - \mu_2)^2/2\sigma_2^2\right] / [2\pi \sigma_1 \sigma_2 (1 - \rho^2)^{1/2}]$,

  where the covariance coefficient $\rho := \text{cov}(X_1, X_2) / (\sigma_1 \sigma_2)$.

• Exercise: Show that if $\rho = 0$ ($X_1$ and $X_2$ are uncorrelated), then these jointly Gaussian random variables are independent!
Joint Gaussian density with $\sigma_1=1$ and $\sigma_2=2$

Uncorrelated case (zero correlation coefficient) at left
Correlation coefficient 0.75 on right

The case of a singular covariance matrix – linearly-dependent centered Gaussian RVs

- The random vector $\mathbf{X} = [X_1 \ X_2 \ ... \ X_n]^T$ will have a singular covariance matrix $\mathbf{C}$ if its component centered random variables are “linearly dependent”, i.e., there are scalars $a_k$, not all zero, such that $\sum_{k=1}^{n} a_k (X_k - \mathbb{E}X_k) = 0$ a.s.
- To see why, note that $\mathbf{C}_a = \mathbb{E}(\mathbf{X} - \mathbb{E}\mathbf{X}) (\mathbf{X} - \mathbb{E}\mathbf{X})^T a = 0$ since $(\mathbf{X} - \mathbb{E}\mathbf{X})^T a = 0$ a.s., where $\mathbf{a} = [a_1 \ ... \ a_n]^T$, and since $\mathbf{a} \neq \mathbf{0}$, we get that $\mathbf{C}$ is singular (with linearly dependent columns).
- Obviously, if a random vector $\mathbf{X} - \mathbb{E}\mathbf{X}$ has linearly dependent components, then they are statistically dependent too since: if $\mathbf{a} \neq \mathbf{0}$ then there is an $a_k \neq 0$ so that $X_i - \mathbb{E}X_i = \sum_{k \neq i} (-a_k/a_i) (X_k - \mathbb{E}X_k)$ a.s.
- **We will describe** a simple procedure to determine a maximal linearly independent subset of $\mathbf{X} = [X_1 \ X_2 \ ... \ X_n]^T$ so that:
  - the multivariate Gaussian PDF can be used for this subset, and
  - the remaining random variables are a.s. linear combinations of those of this subset.
A linear combination of Gaussian RVs is Gaussian distributed

**Theorem:** For any n scalars $a_1, a_2, \ldots, a_n$ and jointly Gaussian $X_1, X_2, \ldots, X_n$, with non-singular covariance matrix $C$, $a_1 X_1 + a_2 X_2 + \ldots + a_n X_n = a^T X$ is Gaussian distributed.

**Proof:**

- If $m = \mathbb{E}X$, the MGF of the linear combination (at $t$) is
  \[
  E \exp(t \ a^T \ X) = \int \exp(t \ a^T \ x) f_X(x) dx \quad (\text{where } x \text{ is an } \mathbb{R}^n - \text{vector})
  \]
  \[
  = [(2\pi)^{-n/2} \det(C)^{-1/2}] \int \exp[t \ a^T \ x - (x - m)^T C^{-1} (x - m)/2 ] \ dx
  \]
- The task now is to “complete the square” in the integrand’s exponent:
  \[
  (x - m)^T C^{-1} (x - m) - 2t a^T x = x^T C^{-1} x - 2(m^T - ta^T C) C^{-1} x + m^T C^{-1} m
  \]
  \[
  = (x - \mu)^T C^{-1} (x - \mu) - \mu^T C^{-1} \mu + m^T C^{-1} m
  \]
  where $\mu := m - tC a$.

A linear combination of Gaussian RVs is Gaussian distributed (cont)

- Thus,
  \[
  E \exp(t \ a^T \ X) = \exp(- \mu^T C^{-1} \mu/2 + m^T C^{-1} m/2)
  \]
  \[
  = \exp(- (m - tC a)^T C^{-1} (m - tC a)/2 + m^T C^{-1} m/2)
  \]
  \[
  = \exp(t \ a^T m + t^2 (a^T C a)/2 )
  \]
- This is the MGF of a Gaussian distributed random variable with mean $a^T m$ and variance $a^T C a$.
- Q.E.D.

**Exercise:** Show that for a random vector $X$ of any distribution with mean $m$ and covariance $C$,
  \[
  E(a^T X) = a^T m \quad \text{and} \quad \text{var}(a^T X) = a^T C a.
  \]
**Computer Exercise A:** Generating independent samples from a distribution

- Generate samples of a uniformly distributed random variable in python (in C, this involves using the srand48 and drand48 functions in double precision)
- Transform them
  - to samples of an exponentially distributed random variable with mean 10
  - to sample of a Bernoulli{5,20} distributed random variable $X$ with $P(X=5)=0.3$ and $P(X=20)=0.7$
- Plot the histogram of the generated samples and compare them against the exponential and Bernoulli, respectively.

**Computer Exercise B:** Generating independent samples of jointly Gaussian random vectors

- Suppose $X \sim N(EX,C)$ for some nonsingular covariance matrix C.
- Since C is positive definite, we can do a Cholesky decomposition to find unique nonsingular lower-triangular $L$ such that $C=LL^T$ and $L^T=L^{-1}$.
- Show that you can write $X = LZ + EX$, where $Z \sim N(0, I)$, i.e., the random variables are uncorrelated (and independent, cf. conditioning section) standard Gaussians.
  - **Note:** The factor $[\text{det}(C)]^{1/2}$ in the PDF for $X$ cancels the Jacobian $\text{det}(L) = [\text{det}(C)]^{1/2}$ when changing variables of (multidimensional) integration to $Z$ from $X = LZ + EX$.
- For the two-dimensional case, write a program that will independently sample $N(EX, C)$ by using the Cholesky decomposition of $C$ and generating independent standard normally distributed samples.
- Generate a two-dimensional scatter plot of the samples for different instances of $EX$ and $C$. 
Computer Exercise C: Cauchy versus Normal—
not all “bell curves” are alike

- Generate $N=10^4$ i.i.d. samples $(X_i, Y_i)_{i=1}^N$ with each $X_i$ and $Y_i$ i.i.d., for
  two cases:
  - standard normal
  - standard Cauchy (density $1/\pi(1+x^2)$).
- Create a scatterplot for each case, i.e., put a dot at $(X_i, Y_i)$ of the
  Cartesian plane.
- In the normal case, you should see a fuzzy disk of radius about 1.
- In the Cauchy case, you should see that the samples tend to hug the
  axes.
- Redo the plots of $(X_i, Y_i)$ conditioning on the sum being large, say
  larger than 30. If that is too high, either lower the threshold or generate
  more points (increase $N$).
- Estimate the conditional distribution of $\max(X,Y)$ given $X+Y = z$.
- Explain what you are finding!

(c) 2013 George Kesidis

Probability for
Electrical and Computer Engineers
Random Variables –
Conditioning and Independence

Professor George Kesidis
kesidis@gmail.com
unless otherwise noted, © 2013 George Kesidis

(c) 2013 George Kesidis
Conditional expected value

- Consider an event A such that \( P(A) > 0 \) and a random variable \( X \).
- The conditional CDF of \( X \) given \( A \) is \( F_{X|A}(z) := P(X \leq z | A) \).
- The conditional expected value of \( X \) given \( A \) is
  \[
  \mu(X|A) = \int_{-\infty}^{\infty} x \, dF_{X|A}(x).
  \]
- In the case of a discretely distributed random variable \( X \), we can use the conditional PMF:
  \[
  \mu(X|A) = \sum_{j=1}^{\infty} a_j P(X = a_j | A) = \sum_{j=1}^{\infty} a_j p_{X|A}(a_j)
  \]
  where the strict range of \( X \) is \( \{a_j\}_{j=1}^{\infty} = R_X \) and \( p_{X|A} \) is the conditional PMF of \( X \) given \( A \).
- Note that some of the conditional PMF terms \( P(X = a_j | A) \) may be zero.
Unconditional CDF of $X \geq 0$ a.s. and conditional CDF given $B = \{X \leq 10\}$

![Graph showing CDFs](image)

Note: $F_X(10|B) \equiv F_{X|B}(10) = 1$

Conditional Expected Value: the PDF

- In a similar way we can define the conditional PDF of a continuously distributed random variable $X$ given the event $A$ (with $P(A) > 0$),
  
  $$f_{X|A}(x) \equiv dF_{X|A}(x) / dx.$$

- So here the conditional expected value is
  
  $$\mu(X|A) = \int_{-\infty}^{\infty} x f_{X|A}(x) \, dx.$$
Communication channel example (SW Ex. 2.6-4) for a slightly more complex setting than BSC

- A signal $Y$ is received as Gaussian distributed but the mean and variance depend on the transmitted signal (symbol) $X \in \{-1, 0, 1\}$, so that $Y \sim N(-1,4)$ with probability $4/7 = P(X = -1)$, $Y \sim N(0,1)$ w.p. $2/7 = P(X = 0)$, and $Y \sim N(1,4)$ w.p. $1/7 = P(X = 1)$.
- The PMF values of $X$ are the prior probabilities for this problem.
- The received noise level (variance of $Y$) depends on the transmitted $X$.
- $Y$ is distributed as a mixture of Gaussians depending on $X$, i.e., $E(Y|X) \sim N(X,V(X))$ where $V$ is given above.
- We can easily compute the distribution of $Y$ by conditioning on $X$, e.g., $P(Y > -1) = \sum_{x \in \{-1, 0, 1\}} P(Y > -1 | X = x)P(X = x)$
  
  $= P(N(1,4) > -1)(1/7) + P(N(0,1) > -1)(2/7) + P(N(-1,4) > -1)(4/7)$
  
  $= ... \quad$ (exercise: evaluate)
- $P(Y > -1 | X = x)$ is a posterior probability.

Example: MAP rule (cont)

- Say we are given the observation at the receiver that $\{Y > -1\}$ and we want to find the most likely transmitted signal $X$ (at the sender).
- That is, we want to select $x \in \{-1, 0, 1\}$ for which $P(X = x | Y > -1)$ is maximal.
- To do this with the given information, recall Bayes’ rule allows us to write in terms of posterior probabilities:
  
  $P(X = x | Y > -1) = P(Y > -1 | X = x)P(X = x)/P(Y > -1)$
  
  where $P(Y > -1)$ is computed by conditioning on $X$ as described above.
- For $x = 1$, this quantity is:
  
  $P(X = 1 | Y > -1) = P(N(1,4) > -1)(1/7)/P(Y > -1)$
  
  $= P(N(1,4) > -1)(1/7) /\{ P(N(1,4) > -1)(1/7) + P(N(0,1) > -1)(2/7) + P(N(-1,4) > -1)(4/7) \}$

- Exercise: Find the maximum a posteriori (MAP) transmitted signal $X$ so that the receiver can best “decode” $\{Y > -1\}$.
- Note how the MAP rule employs the (prior) distribution of $X$. 
Example: The Poisson transform modeling the photo-detector (SW Ex. 2.6-5)

- The number of photons \(X\) striking a photo-detector over a given interval of time is Poisson distributed with parameter \(\mu\), where \(\mu\) is a potentially random quantity depending on the light source.
- Under laser light, \(\mu\) is constant.
- But if the light source is thermal, then \(\mu\) will be exponentially distributed, say with mean \(m := E\mu\).
- So, \(X\) is (discretely) Poisson conditioned on \(\mu\) which is (continuously) exponential.
- To find the distribution of \(X\) in the thermal case, we need to condition on \(\mu\): for \(i \in \{0, 1, 2, 3, \ldots\}\),
  \[
P(X=i) = \int_0^\infty P(X=i \mid \mu = z) f_\mu(z) \, dz
  = \int_0^\infty [e^{-z} z^i / i!] f_\mu(z) \, dz \quad \text{... the Poisson transform of } f_\mu \text{ at } i
  = \int_0^\infty [e^{-z} z^i / i!] (e^{z/m} / m) \, dz \quad \text{... since } \mu \text{ is exp. with mean } m
  
- Note that [... in the integrand is the Poisson PMF with parameter } x.
- Exercise: Evaluate \(P(X=i)\) for the thermal case in terms of \(m\).

**SW Figure 2.6-3** (a) Optical communication system; (b) output current from photo-detector via photo-electric effect.
Conditional expectation

- Consider now two discretely distributed random variables $X$ and $Y$.
- We will define the conditional expectation of $X$ given the random variable $Y$, denoted $E(X|Y)$.
- The quantity $E(X|Y)$ is itself a random variable.
- Indeed, suppose $\{b_j\}_{j=1}^{\infty} = \mathbb{R}$ is the strict range of $Y$.
- For all samples $\omega_j \in \{\omega \in \Omega | Y(\omega) = b_j\} =: B_j$, define the conditional expectation of $X$ given $Y$ as $$E(X|Y)(\omega_j) \equiv \mu(X|B_j),$$
- Note that the conditional expected value on the right-hand side can be denoted $\mu(X|Y = b_j)$.
- That is, the conditional expectation $E(X|Y)$ maps all samples in the event $B_j$ to the conditional expected value $\mu(X|B_j)$, i.e., in this way $X$ has been "smoothened" to produce $E(X|Y)$.
- Therefore, the random variable $E(X|Y)$ is almost surely a function of $Y$.

Conditional expectation – example

- For the purposes of a simple graphical example, suppose that the sample space $\Omega = [0, 1] \subset \mathbb{R}$ (but recall that, in general, $\Omega$ can be a completely abstract space without ordering).
- In the next slide, the previous case of jointly discrete random variables $X$, $Y$ and $E(X|Y)$ are plotted as functions from $\Omega$ to $\mathbb{R}$.
- To further simplify the graph, we assume that these random variables are piecewise constant functions over $\Omega$ and that the probability $P$ is the (Lebesgue) measure corresponding to Euclidean length on $\Omega$.
- Again, $Y^{-1}(b_j) = B_j$
Conditional expectation – example (cont)

The previous example shows that $E(X|Y)$ is a smoothed (less "uncertain") version of $X$ over the events $B_i$.

Also $E(X|Y)$ depends on $Y$ only through the events $B_i \equiv \{Y=b_i\}$, i.e., the extent to which the random variable $Y$ discriminates the samples $\omega \in \Omega$.

Consider another discrete random variable $Z$ with range $R_Z = \{c_j\}_{j=1}^{\infty}$.

Define the events $C_j \equiv \{Z=c_j\}$ for all $j$.

If the collections of events $\{B_j\}_{j=1}^{\infty}$ and $\{C_j\}_{j=1}^{\infty}$ are the same (allowing for differences of probability zero), then

$$E(X|Y) = E(X|Z) \text{ almost surely.}$$

Note that $R_Y$ can be different from $R_Z$, in which case

$$E(X|Y) = h(Y) = E(X|Z) = g(Z) \text{ almost surely, with } g \neq h.$$  

Essentially, knowledge of $Y$ or $Z$ results in the same information because the events $\{B_j\}_{j=1}^{\infty} = \{C_j\}_{j=1}^{\infty}$.

Exercise: Show $E(E(X|Y)) = EX$ for all $X,Y$ for the jointly discrete and jointly continuous cases (smoothing preserves the mean).
Conditional Expectation: the PDF

• Now consider two random variables X and Y which are continuously distributed with joint PDF \( f_{X,Y} \)
• For a fixed (given) \( y \in \mathbb{R} \) such that \( f_Y(y) > 0 \) (i.e., \( y \in \mathbb{R}_Y \)), we can define the conditional density:
  \[
f_{X|Y}(x|y) := \frac{f_{X,Y}(x,y)}{f_Y(y)}
\]
• Exercise: show that for each such \( y \), \( f_{X|Y}(\cdot|y) \) is itself a PDF.
• So, the conditional expected value can be computed as
  \[
  \mu(X|Y=y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx
  \]
• Note that, unlike the discretely distributed case, here the event \( \{Y = y\} \) has zero probability.
• Again note that \( \mu(X|Y = y) \) is a function of \( y \), say \( h(y) \), and that the conditional expectation \( E(X|Y) = h(Y) \).
• We will do an example shortly using the jointly normal PDF.

Conditional Expectation: the PDF - example

• Suppose
  \[
f_{X,Y}(x,y) = \exp(-y-(x-y-y^2/3)^2/2) \cdot u(y)/(2\pi)^{1/2},
\]
  where \( u \) is the unit step.
• First note that the (marginal) distribution of \( Y \) is
  \[
f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dx = e^y \cdot u(y), \text{ i.e., } Y \text{ is exponentially distr'd.}
\]
• Also, for \( y \geq 0 \),
  \[
f_{X|Y}(x|y) = f_{X,Y}(x,y) / f_Y(y) = \exp(-(x-y-y^2/3)^2/2)/(2\pi)^{1/2}
\]
• So, \( E(X|Y) \sim N(Y + Y^2/3, 1) \), i.e., conditioned on \( Y = y \), \( X \) is Gaussian distributed with mean \( y + y^2/3 \) and variance 1.
• Note: The marginal density of \( X \) is not easy to compute.
Conditional expectation is the MMSE estimator

- In general, $E(X \mid Y)$ can be proved to be the measurable function of $Y$ which minimizes the mean-square error (MSE), $E[(X - h(Y))^2]$, among all (measurable) functions $h$.
- As the minimum MSE (MMSE) estimator, $E(X \mid Y)$ is the measurable function of $Y$ that is the “best approximation” of $X$ (given $Y$).

Recall: Independent random variables

- Three random variables $X$, $Y$, $Z$ are said to be (mutually) independent if the events $\{X \in A\}$, $\{Y \in B\}$, $\{Z \in C\}$ are independent for all Borel sets $A, B, C \subseteq \mathbb{R}$.
- A simpler test for independence follows.
- The joint CDF of $X,Y,Z$ is $F_{X,Y,Z}(r,s,t) = P(X \leq r, Y \leq s, Z \leq t)$, $r,s,t \in \mathbb{R}$,
  recall that comma separated events are intersected.
- Since $\{Y \leq \infty\}=\Omega$, one can obtain a marginal CDF from the joint CDF as $F_X(r) = F_{X,Y}(r,\infty)$; also, $F_Y(s) = F_{X,Y}(\infty,s)$.
- By Dynkin’s $\pi/\lambda$-class theorem, $X$ & $Y$ are independent if $F_{X,Y}(r,s) = F_X(r) F_Y(s)$ for all $r,s \in \mathbb{R}$.
- **Exercise:** show that if $X$ and $Y$ are independent, then $g(X)$ and $h(Y)$ are independent for any (measurable) functions $g$ and $h$. 
Conditional expectation under independence

- When $X$ and $Y$ are independent, the conditional distribution of $X$ given $Y$ is just the distribution of $X$ and, therefore, $E(X|Y) = EX$, i.e., the random variable $E(X|Y)$, a function of $Y$, is constant.
- In other words, if $X$ and $Y$ are independent, then knowledge of $Y$ (i.e., given $Y$) does not affect the remaining uncertainty of the random variable $X$.
- For the continuously distributed case,
  
  $$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} \quad \text{for } x \in \mathbb{R}, y \in \mathbb{R} \quad (f_Y(y) > 0)$$
  
  $$= f_X(x) \frac{f_Y(y)}{f_Y(y)} \quad \text{by independence}$$
  
  $$= f_X(x)$$

Independent versus uncorrelated

Example:
- Suppose that $X$ is a continuous random variable uniformly distributed on the interval $[-1,1]$ and that $Y = X^2$.
- Note that $EX^n = \int_{-1}^{1} z^n 0.5 \, dz = 0$ if $n$ is odd.
- $X$ and $Y$ are uncorrelated because $E(XY) = EX^3 = 0 = EX \cdot EY$ (since $EX=0$),
- but $X$ and $Y$ are clearly dependent.
So, uncorrelated does not necessarily imply independent.
If $X_1$ and $X_2$ are independent then:
- they are uncorrelated, i.e., $\text{cov}(X_1, X_2) = 0$, since $EX_1X_2 = \iint x_1x_2 \, dF_{X_1X_2}(x_1,x_2) = \iint x_1x_2 \, dF_{X_1}(x_1)dF_{X_2}(x_2) = EX_1EX_2$
- where the second equality is by independence, and
- $\text{var}(X_1 + X_2) = \text{var}(X_1) + \text{var}(X_2)$. (exercise)
Uncorrelated Gaussians are independent

- Note that if the random variables $X_1, X_2, \ldots, X_n$ are uncorrelated, i.e.,
  \[ E(X_i X_k) = EX_i EX_k \quad \text{for all } i \neq k, \]
  then covariance matrix $C$ is diagonal.
- So it's easy to see that in the jointly Gaussian case
  \[ f_{\mathbf{X}}(\mathbf{x}) \equiv \prod_k f_{X_k}(x_k), \]
  i.e., the joint density separates so the random variables are independent.
- So, in the Gaussian case, uncorrelated does imply independent.
- Again, the converse is generally true, i.e., independence implies uncorrelated.

For jointly Gaussian RVs, the MMSE is linear

- For jointly Gaussian $X$ and $Y$, the conditional density
  \[ f_{Y|X}(y|x) := f_{Y,X}(y,x)/f_X(x) \]
  is also a Gaussian density in $y$ (for a fixed $x$).
- Exercise: Verify by direct computation the conditional expected value of $Y$ given $X=x$ is
  \[ \mu(Y|X=x) = xE(XY)/E(X^2) + E(Y) - E(X) E(XY)/E(X^2). \]
- Thus, the conditional expectation
  \[ E(Y|X) = (X-EX)E(XY)/E(X^2) + EY. \]
  is linear in $X$, i.e.,
  \[ E(Y|X) = aX + b \]
  for scalars $a = E(XY)/E(X^2)$ and $b = E(Y) - E(X) E(XY)/E(X^2)$.
- Exercise: Directly show that these scalars $a$ and $b$ also minimize the MSE $E(Y - (aX+b))^2$.
- So, for jointly Gaussian random variables, the MMSE has a linear form.
Algebra for jointly Gaussian RVs

- Owing to these special properties of jointly Gaussian random variables, linear-algebraic notions apply.
- We can identify the correlation $E(XY)$ as an inner product of $X$ and $Y$ (actually between equivalence classes of RVs equal to each other a.s.), and deem $Y$ and $X$ orthogonal (uncorrelated, independent) if $E(XY)=0$.
- Also the norm of $X$ would be $E(X^2)$.
- To illustrate, we’ll now consider the case of a linearly (and statistically) dependent group of jointly Gaussian random variables.

The case of a singular covariance matrix – linearly-dependent centered Gaussian RVs

- Recall that the random vector $\mathbf{X}=[X_1 \ X_2 \ \ldots \ X_n]^T$ will have a singular covariance matrix $\mathbf{C}$ if its component centered random variables are “linearly dependent”, i.e., there are scalars $a_k$, not all zero, such that $(\mathbf{X}-E\mathbf{X})^T a = \sum_{k=1}^n (X_k - EX_k) a_k = 0$ a.s.
- The following simplification of the Gram-Schmidt procedure, from the theory of linear vector spaces, can identify a maximal linearly independent subset $\mathbf{Y}$ of $\mathbf{X} = [X_1 \ X_2 \ \ldots \ X_n]^T$ so that
  - the multivariate Gaussian PDF can be used for $\mathbf{Y} \sim N(E\mathbf{Y},\mathbf{C}_Y)$ with nonsingular $\mathbf{C}_Y$.
  - the remaining random variables of $\mathbf{X}$ are a.s. linear combinations of the elements of $\mathbf{Y}$:
  1. Include $X_1 \neq EX_1$ a.s. in the (initially empty) set of random variables $\mathbf{Y}$.
  2. For each $k>1$, find the MMSE estimator of $X_k - EX_k$ given $\mathbf{Y} - E\mathbf{Y}$, and if the MMSE $\neq 0$ then include $X_k$ in $\mathbf{Y}$. 

(c) 2013 George Kesidis
The case of a singular covariance matrix (cont)

- Recall that for jointly Gaussian random variables, the MMSE estimator is linear.
- So, to find the MMSE estimator of $X_k - EX_k$ given $Y - EY$, the first order optimality conditions of the MSE
  $$E[(X_k - EX_k) - (Y - EY)^T a]^2$$
  are
  $$C_{Y \alpha} = E[(Y - EY)(X_k - EX_k)] =: C_{Y, X_k}$$
  where $C_{Y, X_k}$ is the column vector of covariances between $Y$ and $X_k$.
- By solving for the scalars $a$, we get that the MMSE estimator of $X_k - EX_k$ given $Y - EY$ is
  $$(Y - EY)^T C_{Y}^{-1} C_{Y, X_k}$$
- So, the MMSE is
  $$E[(X_k - EX_k) - (Y - EY)^T C_{Y}^{-1} C_{Y, X_k}]^2 = \text{var}(X_k) - C_{Y, X_k}^T C_{Y}^{-1} C_{Y, X_k}$$

Sums of independent RVs – convolution of PDFs (or PMFs)

- Consider a sum of mutually independent continuous random variables.
- Our objective is to find the PDF of the sum armed with the PDF of the component random variables.
- To this end, consider two independent random variables $X_1$ and $X_2$ with PDFs $f_1$ and $f_2$ respectively; so, $f_{X_1,X_2}(z) \equiv f_1(x_1)f_2(x_2)$.
- Thus, the CDF of the sum $X_1 + X_2$ is
  $$F(z) = P(X_1 + X_2 \leq z) = \int_{-\infty}^{z} f_1(x_1) dx_1 = \int_{-\infty}^{z} f_2(z-x_1) f_1(x_1) dx_1$$
  - Exchanging the integral on the right-hand side with a derivative with respect to $z$, we show that PDF of $X_1 + X_2$ is
    $$f(z) = F'(z) = \int_{-\infty}^{z} f_2(z-x_1) f_1(x_1) dx_1, \ z \in \mathbb{R}.$$ 
- Thus, $f$ is the convolution of $f_1$ and $f_2$, denoted $f = f_1 * f_2$, i.e.,
  $$(f_1 * f_2)(z) = \int_{-\infty}^{z} f_2(z-x_1) f_1(x_1) dx_1, \ z \in \mathbb{R}.$$
Note: convolution notation

- It is common to see a convolution operation written as $f_1(z) * f_2(z)$, but you should never do this.
- A convolution operation takes two whole signals (or functions), say $f_1$ and $f_2$, and produces another signal, their convolution $f_1 * f_2$.
- The convolution signal can be evaluated at a point $z$ in its domain, i.e., $(f_1 * f_2)(z)$.
- The notation $f_1(z) * f_2(z)$ seems to imply that the convolution at $z$ is produced from a single sample of the two functions being convolved, i.e., $f_1(z)$ and $f_2(z)$, which is simply not the case.

Sums of independent RVs - MGFs

- In this context, moment generating functions can be used to simplify calculations.
- Let the MGF of $X_i$ be
  $$m_i(\theta) = E e^{\theta X_i} = \int_{-\infty}^{\infty} f_i(x) e^{\theta x} dx.$$  
  Note that $m_i$ is basically the (bilateral) Laplace transform of $f_i$.
- The MGF of $X_1 + X_2$ is
  $$m(\theta) = E e^{\theta (X_1 + X_2)} = E e^{\theta X_1} e^{\theta X_2} = m_1(\theta)m_2(\theta),$$  
  where the last equation holds because of assumed independence.
- So, convolution of PDFs corresponds to simple multiplication of MGFs, which, in turn, corresponds to addition of independent random variables.
Exercise: The Poisson distribution

- Recall that we showed using MGFs that a linear combination of Gaussian random variables is Gaussian distributed.
- **Exercise:** Use MGFs to show that the sum of independent Poisson distributed random variables with parameters $a$ and $b$ is Poisson distributed with parameter $a+b$.

Example: The gamma distribution

- As an example, suppose $X_1$ and $X_2$ are independent and both exponentially distributed with parameter $\lambda>0$.
- The PDF of $X_1 + X_2$ is $f$, where $f(z) = 0$ for $z < 0$ and, for $z \geq 0$,
  \[ f(z) = \int_{-\infty}^{\infty} f_1(x_1) f_2(z-x_1) dx_1 = \lambda^2 e^{\lambda z} \int_0^z dx_1 = \lambda^2 z e^{\lambda z} \]
- The MGF of $X_1 + X_2$ is
  \[ m(\theta) = \left[ \frac{\lambda}{\lambda - \theta} \right]^2, \]
  which is consistent with the PDF $f$ of the sum as there is a one-to-one relationship between PDFs and MGFs of nonnegative random variables.
- In this case, the MGF is a unilateral Laplace transform of the PDF:
  \[ m(\theta) = \int_0^\infty f(z) e^{\theta z} \, dz. \]
Example: The gamma distribution (cont)

• Indeed, for $n$ i.i.d. $\text{exp}(\lambda)$ RVs, the MGF of $\sum_{i=1}^{n} X_i$ is easily computed as $m(\theta) = \left[ \frac{\lambda}{\lambda - \theta} \right]^n$.
• This is the MGF of a gamma distributed random variable with parameters $\lambda > 0$ and $n \in \mathbb{Z}^+$, where PDF for $z \geq 0$ is $f_n(z) = \lambda^n z^{n-1} e^{-\lambda z} / (n-1)!$
• When the parameter $r \in \mathbb{R}^+$ is a positive integer ($r=n$), the (normalizing) gamma function $\Gamma(r) := \int_0^\infty \lambda^n z^{n-1} e^{-\lambda z} dz = \int_0^\infty y^{r-1} e^{-y} dy = (r-1)!$
  (just integrate by parts $r-1$ times); in this case, the gamma distribution is sometimes called the Erlang distribution, particularly in the queuing literature.

Chi-square distribution

Suppose $X_1, X_2, X_3, ...$ are i.i.d. standard-normally distributed random variables and define $Z_n := X_1^2 + X_2^2 + ... + X_n^2$

**Theorem:** $Z_n$ has chi-square ($\chi^2$) distribution with $n$ degrees of freedom, i.e., the PDF of $Z_n$ is $f_n(z) = z^{1+n/2} e^{-z/2} 2^{-n/2} / \Gamma(n/2)$ for $z \geq 0$, and $f_n(z) = 0$ for $z < 0$.

**Note:** The chi-square distribution has important applications to statistical confidence and has a similar form to that of the gamma distribution.

**Proof:**
• First note that for $n=1$, $X \sim \mathcal{N}(0,1)$ and $z \geq 0$:
  $P(X^2 \leq z) = P(-\sqrt{z} \leq X \leq \sqrt{z}) = F_X(\sqrt{z}) - F_X(-\sqrt{z}) + P(X=-\sqrt{z})$, where $P(-\sqrt{z} = X) = 0$ since $X$ is continuously distributed.
• So differentiating w.r.t. $z$, we get that $f_1(z) = \left[ f_X(\sqrt{z}) + f_X(-\sqrt{z}) \right] / (2z^{1/2}) = f_X(\sqrt{z}) / z^{1/2} = e^{-z/2} / (2\pi)^{1/2}$
• Exercise: Check that $\Gamma(1/2) = \sqrt{\pi} = \pi^{1/2}$. Hint: Change the dummy variable of integration $y = v^2 / 2$ in the definition $\Gamma$ (previous slide).
Chi-square distribution (cont)

• Assume $Z_n$ has PDF $f_n$ for $n=k \geq 1$.
• To complete an inductive proof of the theorem, we need to show $Z_{k+1}$ has PDF $f_{k+1}$.
• To do this, note that since the $X$'s are independent, $Z_k$ is independent of $X_{k+1}$.
• Thus, $Z_{k+1}$ has PDF $f_k \ast f_1$.
• **Exercise:** perform the convolution to show $f_{k+1} = f_k \ast f_1$.
• Q.E.D.

Again, a direct proof is more easily obtained by MGFs:

• Establish $Z_1$ has PDF $f_1$ as in the first step of the inductive proof.
• Let the MGF of $Z_1$ be $m_1$, i.e., $m_1(t) := \int_{-\infty}^{\infty} f_1(z) e^{tz} \, dz$ (= ... exercise)
• By independence of the $X$'s, the MGF of $Z_n$, $m_n = (m_1)^n$
• Verify that the MGF $m_n$ corresponds to the PDF $f_n$ given in theorem stmt

The Chi-square probability density function for $n = 2$ (solid), $n = 4$ (dashed), and $n = 10$ (stars).
Remark: Bandpass noise

- Suppose at frequency \( w \), a communication channel carries in-phase (say cosine) and quadrature-phase (sine) "bandpass" noise components:
  \[ Z(t) = X \cos(\omega t) + Y \sin(\omega t) \]

- Exercise: If \( X \) and \( Y \) are i.i.d. \( \mathcal{N}(0, \sigma^2) \), i.e., in-phase and quadrature-phase noise are uncorrelated, show that the amplitude of \( Z \) is Raleigh distributed with parameter \( \sigma \), i.e., that \( Z \) has PDF
  \[ f(x) = \frac{x}{\sigma^2} \exp(-x^2/\sigma^2) \text{ for } x \geq 0, \text{ and } f(x) = 0 \text{ for } x < 0. \]

- Note that the square of the amplitude of \( Z \), i.e., \( X^2 + Y^2 \), is chi-square distributed when \( \sigma = 1 \), otherwise gamma distributed.

- Note: If after transmission through a channel, the deterministic sinusoidal signal \( S(t) = V \cos(\omega t + \theta) \), with \( V \geq 0 \), is received with this bandpass noise added to it, i.e., \( R(t) = S(t) + Z(t) \), then the amplitude of \( R \) is Rice distributed (Rician), i.e., \( (A^2 + B^2)^{0.5} \) where independent \( A = X + V \cos(\theta) \sim \mathcal{N}(V \cos(\theta), \sigma^2) \) and \( B = Y + V \sin(\theta) \sim \mathcal{N}(V \sin(\theta), \sigma^2) \)

The Rayleigh, exponential, and uniform PDFs

![Rayleigh, exponential, and uniform PDFs](image)
Useful Inequalities: Boole’s inequality

- Clearly, if the event $A_1 \subset A_2$, then
  $$P(A_1) \leq P(A_2) = P(A_1) + P(A_2 \setminus A_1).$$
- For any collection of events $A_1, A_2, \ldots, A_n$, Boole’s inequality holds:
  $$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$
- Note that when the $A_i$ are disjoint, equality holds simply by the additivity property of a probability measure $P$. 
Useful Inequalities: Markov’s inequality

• If two random variables $X$ and $Y$ are such that $P(X \geq Y) = 1$ (i.e., $X - Y \geq 0$ almost surely), then it’s easy to show that $EX \geq EY$.

• Consider a random variable $X$ with $E|X| < \infty$ and a real number $x > 0$.

• Since $|X| \geq |X|1\{|X| \geq x\} \geq x1\{|X| \geq x\}$ almost surely, we arrive at Markov’s inequality:

$$E|X| \geq Ex1\{|X| \geq x\} = xE1\{|X| \geq x\} = xP(|X| \geq x).$$

• An alternative explanation for continuously distributed random variables $X$ (with PDF $f$) is, for $x \geq 0$,

$$E|X| = \int_{-\infty}^{\infty} z|f(z)|dz$$

$$\geq \int_{-\infty}^{\infty} z|f(z)|1\{|z| \geq x\}dz$$

$$= \int_{-\infty}^{-x} (-z)f(z)dz + \int_{x}^{\infty} zf(z)dz$$

$$\geq \int_{-\infty}^{-x} xf(z)dz + \int_{x}^{\infty} zf(z)dz$$

$$= x[\int_{-\infty}^{-x} f(z)dz + \int_{x}^{\infty} f(z)dz]$$

$$= x[P(X \leq -x) + P(X \geq x)]$$

$$= xP(|X| \geq x)$$

Useful Inequalities: Markov’s inequality (cont)

• Writing Markov’s inequality as

$$P(|X| \geq x) \leq x^{-1}E|X|$$

• So, Markov’s inequality allows us to bound the “tail” of the distribution of a random variable by simply using its average value.

• Such tail probabilities are often very small in practice (corresponding to rare events corresponding “errors”).

• Compared to such tail probabilities, the average value of a random variable is often much easier to exactly compute given its distribution or much easier to accurately estimate through sampling.

• Often tail probabilities are expressed in terms of deviations from the mean and, to this end, we can replace $X$ by $X - EX$ in Markov’s inequality.
Chebyshev’s Inequality

• Now take \( x = \varepsilon^2 \), where \( \varepsilon > 0 \), and argue Markov’s inequality with \( (X - EX)^2 \) in place of \( |X| \) to get Chebyshev’s inequality, i.e.,

\[
\text{var}(X) := E[(X - EX)^2] \geq \varepsilon^2 P(|X - EX| \geq \varepsilon),
\]

so that

\[
P(|X - EX| \geq \varepsilon) \leq \varepsilon^{-2} \text{var}(X).
\]

• Exercise: If \( \sigma = [\text{var}(X)]^{0.5} \) is the standard deviation of \( X \), then show that Chebyshev’s inequality can be written as

\[
P(|X-EX|>z\sigma) \leq z^{-2} ;
\]

note how the inequality is vacuous for \( z<1 \).

Chebyshev’s Inequality - example

• Suppose \( X \sim \text{Gamma}(m) \) for integer \( m \geq 0 \) (i.e., Erlang) with PDF

\[
f(z) = z^m e^{-z}/m! \quad \text{for real } z>0, \text{ and } =0 \text{ else}.
\]

• That is, \( X \sim \) a sum of \( m+1 \) i.i.d. \( \exp(1) \) random variables.

• We now show that \( P(0<X<2(m+1)) > m/(m+1) \):

\[
P(0<X<2(m+1)) = 1 - P(X \geq 2(m+1))
\]

\[
= 1 - P(X-(m+1) \geq m+1) \quad \text{where } EX = m+1
\]

\[
\geq 1 - P(|X-(m+1)| \geq m+1) \quad \text{since } |Y| \geq Y
\]

\[
\geq 1 - \text{var}(X)/(m+1)^2 \quad \text{Chebyshev’s inequ.}
\]

\[
= 1 - (m+1)/(m+1)^2
\]

• See Rozanov p. 80, prob. 3.

• Exercise: Verify that \( EX = m+1 \) and \( \text{var}(X) = (m+1)m \), by respectively integrating-by-parts just once and twice, and compute the error in the Chebyshev inequality for specific values of \( m \).

• Note that using Markov’s inequality is less informative:

\[
P(0<X<2(m+1)) = 1 - P(X > 2(m+1)) \geq 1-(m+1)/(2(m+1)) = 0.5
\]
Cramer Inequality or Chernoff Bound

- Note, for all θ>0, the event \{X≥x\}={e^{θX}≥e^{θx}}.
- Arguing as for Markov’s inequality gives Cramer’s inequality:
  \[ Ee^{θX} ≥ e^{θx}P(X ≥ x) \]
  \[ ⇒ P(X ≥ x) ≤ \exp(−[xθ − \log Ee^{θX}]) \]
  \[ ⇒ P(X ≥ x) ≤ \exp(−\max_{θ>0} [xθ − \log Ee^{θX}]) \]
  where we have simply sharpened the inequality by taking the maximum over the free parameter θ.
- The result is the Legendre transform of the log-MGF of X in the exponent of Cramer’s inequality (a.k.a. Chernoff bound).

Chernoff bound – example

- For n i.i.d. Bernoulli\{0,1\} random variables with p(1)=q <0.5.
- Let S be the sum of the n random variables and X be distributed as one of them.
- We want to compute P(S≥n/2), the probability that the majority vote against their bias (q<0.5).
- We can compute this using the fact that S is binomially distributed.
- For a simpler approximation, we can use the Chernoff bound with the equation \(Ee^{θX}=(1-q+qe^{θ})^n\) where the first equality follows from the i.i.d. assumption on the Bernoulli samples.
- The exponent in the Chernoff bound is therefore
  \[ \max_{θ>0} [θn/2 − \log(Ee^{θX})^n] = n \max_{θ>0} [θ/2 − \log(1-q+qe^{θ})] \]
- Exercise: Find and substitute in the maximizing θ to get that
  \( P(S≥n/2) ≤ (-1+1/q)\exp(-n(0.5-\log(2q))) \)
- Based on this bound, how large should n be if we want to be able to assert that \( P(S≥n/2) < ε \) for some positive \( ε≪1 \).
Cauchy-Bunyakovsky-Schwarz inequality

- The CBS (commonly just Cauchy-Schwarz) inequality states that
  \[ E(XY) \leq [E(X^2)]^{0.5}[E(Y^2)]^{0.5} \]
  for all random variables X and Y, with the inequality strict
  whenever Y \neq 0 \neq X a.s. and there is no constant c such that X \neq cY a.s.
- This inequality is an immediate consequence of the fact that
  \[ E(X/[E(X^2)]^{0.5} - Y/[E(Y^2)]^{0.5})^2 \geq 0 \]
  whenever X \neq 0 and Y \neq 0 almost surely.
- Also, note that if we take Y = 1 almost surely, Cauchy-Schwarz simply
  states that the variance of a random variable X is not negative, i.e., that
  \[ E(X^2) - (EX)^2 \geq 0. \]

Cauchy-Schwarz inequality - application

**Theorem:** For a random variable X with finite mean and variance,
\[ P(X > z \mu) \geq (1 - z^2)(EX)^2 / E(X^2) \]
for \(0 \leq z \leq 1.\)

**Proof:**
- Recall that \( P(X > a) = E[1 \{X > a\}]. \)
- Let \( \mu = EX. \)
- By the Cauchy-Schwarz inequality,
  \[ |E(X1\{X > z\mu\})|^2 \leq E(X^2) \ P(X > z\mu). \]
- Also, \( E(X1\{X > z\mu\}) = \mu - E(X1\{X \leq z\mu\}) \geq \mu - z\mu. \)
- Combine the last two inequalities to get the desired result.
- Q.E.D.

Exercise: Fill in the details of the steps of this proof.
Jensen’s Inequality

- That variance is always non-negative is also an immediate consequence of Jensen’s inequality.
- A real-valued function $g$ on $\mathbb{R}$ is said to be convex if
  $$g(px + (1 - p)y) \leq pg(x) + (1 - p)g(y)$$
  for any $x, y \in \mathbb{R}$ and any real fraction $p \in [0, 1]$.
- If the inequality is reversed in this definition, the function $g$ would be concave.
- For any convex function $g$ and random variable $X$, we have Jensen’s inequality:
  $$g(\mathbb{E}X) \leq \mathbb{E}(g(X)).$$
- For a discrete random variable $X$ with PMF $p$ and strict range $\mathbb{R}$, Jensen’s inequality for convex $g$ is basically the definition of convexity,
  $$\sum_{x \in \mathbb{R}} g(x) p(x) \leq g(\sum_{x \in \mathbb{R}} x p(x)).$$

Recall: Probability through statistics

- Using atmospheric noise, random (i.e., equally likely, uniformly random) and independent samples from $\Omega=\{1,2,3,...,100\}$ are tabulated at the web site random.org
- Let $T_k(n)$ be how many times the number $k$ appears in the first $n$ samples of the random.org table,
  $$T_k(n) = \sum_{i=1}^n 1(X_i=k),$$
  where $X_i$ is the $i$th sample.
- The following two graphs plot a single sample path of $T_5(n)/n$ versus $n$ and $T_{23}(n)/n$ versus $n$, respectively, each of which are seen to be converging to $0.01 = 1/100$.
- Later, to prove the weak law of large numbers, we will establish that $T_k(n)$ is a random variable with
  - mean (average value, expectation) $\mathbb{E} T_k(n)/n = 1/100$,  
  - variance $\to 0$ as $n \to \infty$, and
- Also, we will understand $T_5(n)/n$ or $T_7(n)$ as discrete-time stochastic processes in $n$.  

(c) 2013 George Kesidis
$T_5(n)/n$ sample path from random.org

$T_{23}(n)/n$ sample path from random.org
The Weak Law of Large Numbers

- Concretely, a law of large numbers ties statistics to probability.
- Suppose we have an i.i.d. sequence of random variables $X_1, X_2, X_3, \ldots$.
- Also suppose that the common distribution has finite variance, i.e., $\sigma^2 := \text{var}(X) := E(X - EX)^2 < \infty$, where $X \sim X_i$.
- Finally, suppose that the mean exists and is finite, i.e., $-\infty < \mu := EX < \infty$.
- Define the sum $S_n = X_1 + X_2 + \cdots + X_n$ for $n \geq 1$.

Exercise: Prove that $E S_n = n \mu$ and $\text{var}(S_n) = n \sigma^2$.

Weak law of large numbers

- The quantity $M_n := S_n/n$ is called the empirical or sample mean of the sequence $X_k$ after $n$ samples and is an unbiased estimate of $\mu = EX$, i.e., $E M_n = \mu$.
- Also, because of the following weak LLN, $M_n$ is said to be a weakly consistent estimator of $\mu$.

**Theorem:** For all $\varepsilon > 0$,
$$\lim_{n \to \infty} P(|M_n - \mu| \geq \varepsilon) = 0.$$  

**Proof:**
- First note that $\text{var}(S_n) = n \sigma^2$ implies $\text{var}(M_n) = \text{var}(S_n/n) = \text{var}(S_n)/n^2 = \sigma^2/n$.
- So, by Chebyshev's inequality,
$$0 \leq P(|M_n - \mu| \geq \varepsilon) \leq \text{var}(M_n)/\varepsilon^2 = \sigma^2/(n\varepsilon^2) \to 0 \text{ as } n \to \infty.$$  
- Q.E.D.
Weak law of large numbers (cont)

• Suppose the distribution of $X$ is not known and $\mu = \mathbb{E}X$ (also not known) and is to be approximated by the empirical mean $M_n$.

• Note how Chebyshev’s inequality gives us a measure of how “confident” we are that the empirical mean $M_n$ differs from the true mean $\mu$ after $n$ samples $X_1, X_2, \ldots, X_n$.

• But Chebyshev’s bound requires knowledge of the true variance $\sigma^2 = \text{var}(X)$, which could also be approximated by the sample variance (itself using $M_n$).

• In the following, we approximate $P(|M_n - \mu| \geq \varepsilon)$ using a Gaussian distribution, as justified by the Central Limit Theorem (CLT).

Statement of the SLLN

• The strong LLN asserts that, if $\mathbb{E}|X| < \infty$, then
  
  $P(\lim_{n \to \infty} \frac{S_n}{n} = \mu) = 1$.

• In other words, $S_n/n \to \mu$ almost surely (a.s.).

• So, $S_n/n$ is said to be a strongly consistent estimator of $\mu$.

• The first SLLN is due to Kolmogorov.

• Almost sure convergence implies convergence in probability (WLLN), which in turn implies convergence in distribution (CLT).
Recall the Gaussian/Normal distribution

- Assume \( X \) is Gaussian distributed with mean \( \mu \) and variance \( \sigma^2 \), i.e., \( X \sim N(\mu, \sigma^2) \).

- The PDF of \( X \) is
  \[
  p_X(x) = \exp\left[-(x-\mu)^2/(2\sigma^2)\right]/\sqrt{2\pi}, \quad x \in \mathbb{R}
  \]

- Recall that the MGF of \( X \) is
  \[
  m_X(\theta) := E e^{\theta X} = \exp(\mu \theta + \sigma^2 \theta^2/2)
  \]

- The “standard” normal distribution is \( N(0,1) \), i.e., with mean zero and variance one.
Lemma: There is a constant $\beta > 0$ such that $\beta = \lim_{n \to \infty} n!e^n/n^{n+0.5}$

Proof:
• Define $B(n) = n!e^n/(n^{n+0.5})$.
• $\log B(n) = 1 + \sum_{j=2}^{n} \{\log B(j) - \log B(j-1)\}$ where $1 = \log B(1)$.
• By Taylor’s theorem, $\log(1-x) = -x - x^2/2 - x^3/3 + o(x^3)$
  where $\lim_{y \to 0} o(y)/y = 0$.
• So, $\log B(j) - \log B(j-1) = 1 + (j-1/2)\log(1-1/j) = -1/(12j^2) + o(1/j^2)$
• Since $j^{-2}$ is summable ($\int_1^{\infty} x^{-2} \, dx < \infty$), $B(n)$ converges to a finite $\beta$.
• Q.E.D.
de Moivre-Laplace Central Limit Theorem (CLT)

- Consider a sequence of independent and identically distributed (i.i.d.) Bernoulli random variables $X_1, X_2, \ldots$, where
  - $p := P(X_i = 1)$ and $q := 1 - p = P(X_i = 0)$.
- Define the sum $S_n = X_1 + X_2 + \ldots + X_n$.
- $S_n$ is binomially distributed with parameters $(n, p)$, i.e.,
  - $P(S_n = k) = \binom{n}{k} p^k q^{n-k}$ for all $k \in \{0,1,2,\ldots,n\}$.
  - $E S_n = np$ and $\text{var}(S_n) = npq$.
- Thus $Y_n := (S_n - np)/(nqp)^{0.5}$ is centered ($EY_n = 0$) by the linearity of expectation.
- Note that if $X$ has variance $\sigma^2$ then for all scalars $c$, $cX$ has variance $c^2 \sigma^2$, again by linearity of expectation.
- Thus, $Y_n$ also has unit variance, i.e., $\text{var}(Y_n) = 1$ for all $n$.
- Exercise: Recall and prove that the variance of the sum of a group of independent random variables is the sum of their variances.


de Moivre-Laplace CLT statement

- **Theorem (de Moivre-Laplace CLT):** If $X_i$ are i.i.d. Bernoulli random variables, then $Y_n$ defined above converges in distribution to a standard normal (Gaussian), i.e.,
  $$\lim_{n \to \infty} P(Y_n > y) = \Phi(y) = \int_y^{\infty} (2\pi)^{-0.5} \exp(-x^2/2) \, dx$$
- This CLT was a major milestone in 18th century mathematics.
de Moivre-LaPlace CLT - Proof

• Let \( z(n) := (np)^{0.5} \)

• \( P(a < Y_n := (S_n - np)/z(n) \leq b) = \sum_{az(n) < k \leq bz(n)} C(n, k)p^kq^{n-k} \)
  \( = \sum_{az(n) < k' \leq bz(n)} C(n, k' + np)p^{k'+np}q^{n-k'} \)
  where the sums (of the binomial PMF) are over integers \( k, k' \).

• Now use de Moivre’s formula to uniformly approximate \( C(n, k' + np) \) over \( k' \) as \( n \to \infty \) (where \( \sim \) means equal as \( n \to \infty \)):
  \( P(a < Y_n \leq b) \sim (1/\beta z(n)) \sum_{az(n) < k' \leq bz(n)} (1 + k'/np)^{-k' - np}(1 - k'/nq)^{k' + nq} \)
  \( \sim (1/\beta z(n)) \sum_{az(n) < k' \leq bz(n)} \exp(-k'/z(n))^2 \)
  \( \sim (1/\beta) \int_{a}^{b} \exp(-x^2/2) \, dx \)
  \( \sim \frac{1}{\beta} \left( \Phi(b) - \Phi(a) \right) \)
  where the first step is \( \lim_{v \to \infty} (1+1/v)^v = e \), the second step is \( \log(1-x) = -x - x^2/2 + o(x^2) \), and the third step is the Riemann integral with \( x \)-axis partition-width \( z(n) \).

• Taking \(-a, b \to \infty\), gives Wallis' identity for the value of
  \( \beta = \int_{-\infty}^{\infty} \exp(-x^2/2) \, dx = (2\pi)^{0.5} \)
  Q.E.D.

(c) 2013 George Kesidis

Aside: direct proof of Wallis’ identity

**Theorem:** \( \int_{-\infty}^{\infty} \exp(-x^2/2) \, dx = (2\pi)^{0.5} \)

**Proof:**

• Let \( J = \int_{-\infty}^{\infty} \exp(-x^2/2) \, dx \) and note that

• \( J^2 = \int_{-\infty}^{\infty} \exp(-x^2/2) \, dx \int_{-\infty}^{\infty} \exp(-y^2/2) \, dy \)
  \( = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \exp(-[x^2+y^2]/2) \, dx \, dy \)
  \( = \int_{0}^{2\pi} \int_{0}^{\infty} \exp(-r^2/2) r \, dr \, d\theta \) ... change to polar coords
  \( = -2\pi \exp(-r^2/2) \bigg|_{0}^{\infty} \)
  \( = 2\pi \)
  Q.E.D.

(c) 2013 George Kesidis
Stirling’s Formula

• de Moivre’s formula with Wallis’ identity give the famous result of Stirling:
  \[(2\pi)^{0.5} = \lim_{n \to \infty} n! e^n / n^{n+0.5}\]

• For finite \( n \), Stirling’s “formula” is an often-used approximation of \( n! \):
  \[n! \approx (2\pi)^{0.5} n^{n+0.5} e^{-n}\]

• In the following, we state a more general sequential CLT for i.i.d. random variables accommodating a very wide range of possible distributions, i.e., not just Bernoulli.

More general CLT for sequences of independent random variables

• **Theorem:** If \( X_1, X_2, \ldots \) are independent and identically distributed (i.i.d.) with \( E|X_1| < \infty, 0 < \sigma^2 := \text{var}(X_1) < \infty \), then
  \[(S_n - n\mu) / (\sigma\sqrt{n}) \to^d N(0,1),\]
i.e., converges in distribution to a standard normal, where
  \[\mu := E X_1 \text{ and } S_n := X_1 + X_2 + \ldots + X_n.\]

• A proof of this theorem also involves the use of Taylor’s theorem to the second order.

• In the 20th century, the Lindeberg-Feller CLT for sequences of independent, though not necessarily identically distributed, random variables whose (marginal) CDFs satisfy Lindeberg’s (necessary and sufficient) condition.

• We will see how we can approximate \( S_n \) as \( N(n\mu, n\sigma^2) \)-distributed, even for surprisingly small \( n \), when we discuss statistical confidence (for very small \( n \), cf. the “Student’s t” distribution).
Statistical confidence: The empirical mean

- Suppose we wish to determine the mean \( \mu \) of an unknown distribution, e.g., governing the outcome of an election involving a very large number of voters where \( \mu = P(\text{vote for candidate C}) = E(\text{vote C}) \).
- To this end, suppose that a relatively small number \( n \) of i.i.d. samples \( X_1, X_2, X_3, ..., X_n \) are produced (by polling, measurement, simulation, etc.), each distributed according to this unknown distribution.
- Let \( \sigma > 0 \) be the standard deviation of the \( X_k \).
- After \( n \) samples, the sample mean (or empirical mean) is the random variable
  \[ M_n = \frac{1}{n} \sum_{k=1}^{n} X_k = \frac{S_n}{n} . \]
- Note that \( \mu = E \left[ M_n \right] \).
- To develop some measure of how confident we are that \( \mu \approx M_n \) after just \( n \) samples, we will employ the central limit theorem for the sequence \( \{X_k\} \):
  \[ (M_n - \mu) \frac{n^{0.5}}{\sigma} = \left( \frac{S_n}{\sigma \sqrt{n}} - \frac{n \mu}{\sigma} \right) \approx_d N(0,1) , \]
  i.e., is approximately distributed as standard Gaussian.

Statistical confidence: The (biased) sample variance

- Typically, in practice, the standard deviation \( \sigma \) is also not known and must also be estimated from the samples \( X_k \).
- The sample variance is the random variable
  \[ V_n := \frac{1}{n-1} \sum_{k=1}^{n} (X_k - M_n)^2 = \left( \frac{1}{n} \sum_{k=1}^{n} X_k^2 \right) - M_n^2 \]
- The implicit assumption is that a more general form of central limit theorem holds when the sample standard deviation, \( V_n^{0.5} \), is used instead of the true, but unknown, standard deviation \( \sigma \) above.
- That is, we assume \( (M_n - \mu) \frac{n^{0.5}}{V_n^{0.5}} \) is approximately \( N(0,1) \) distributed.
The 95% confidence interval

- So, according to the $N(0,1)$ distribution
  \[ P\left(|(M_n - \mu) n^{0.5} / V_n^{0.5}| \leq 2\right) \approx 0.95. \]
- That is, 95% of the mass of a normal density is within two standard deviations of its mean, cf. the graph on the next slide.
- The inequality $|(M_n - \mu) n^{0.5} / V_n^{0.5}| \leq 2$ is equivalent to
  \[ M_n - 2V_n^{0.5} / n^{0.5} \leq \mu \leq M_n + 2V_n^{0.5} / n^{0.5} \]
- Note how the true mean $\mu$ is bounded above and below by terms directly derived from the data $X_1, X_2, \ldots, X_n$
- So, we say the 95% confidence interval for $\mu$ is taken to be
  \[ [M_n - 2V_n^{0.5} / n^{0.5}, M_n + 2V_n^{0.5} / n^{0.5}] \]
- Exercise: Find the 68% confidence interval. Hint: 68% of the mass of a normal density is within one standard deviation of its mean.

95% probability within two std dev of $N(0,1)$ PDF
Unbiased estimate of sample variance

- It turns out that the sample variance defined above is a consistent (cf. the weak law of large numbers),
\[ \lim_{n \to \infty} V_n = \sigma^2 \]
- But \( V_n \) is a biased estimator for \( \sigma^2 \), i.e., for finite number of samples \( n \),
\[ EV_n = \sigma^2 (n-1)/n \neq \sigma^2 \]
- So, particularly for a small number of samples \( n \), use
\[ U_n := V_n \frac{n}{n-1} = (n-1)^{-1} \sum_{k=1}^{n} (X_k - M_n)^2 \]
instead of \( V_n \) in the 95% confidence interval above.
- That is, \( U_n \) is both a consistent and unbiased estimator of \( \sigma^2 := \text{var}(X_k) \).

Recursive formulas for statistical confidence

- The following simple recursive formulas in the number of observed samples for the sample mean and sample standard deviation are computationally convenient (to update estimates as new samples are observed/generated).
- For the sample mean, we clearly have
\[ M_n = [(n-1)M_{n-1} + X_n]/n. \]
- Also, the sample variance satisfies
\[ V_n = V_{n-1} + [-V_{n-1} + (X_n - M_n)^2] / n. \]
- So for the unbiased sample variance \( U_n := V_n \frac{n}{n-1} \),
\[ U_n = [(n-2) U_{n-1} + (X_n - M_n)^2] / (n-1). \]
- Exercise: Verify these recursions are correct by induction.
Stopping criterion for statistical confidence – relative error

- In order to arrive at a criterion to terminate a simulation, one can define the relative error
  \[ \xi_n := \frac{\text{var}(M_n)^{0.5}}{|M_n|} = \frac{U_n^{0.5}}{|M_n|^{0.5}}, \]
  where the second equality is by assumed independence of the data samples \( X_i \).
- Here we have assumed the sample mean \( \text{EX}_i = \mu \neq 0 \) and finite \( \text{var}(X_i) = \sigma^2 < \infty. \)
- Data generation may be terminated when the relative error reaches, say, 0.1, i.e., the sample standard deviation is 10% of the sample mean.

Stopping criterion for statistical confidence – relative error (cont)

- Note that we can express the statement of the confidence interval as
  \[ P(\mu/M_n \in [1-2\xi_n, 1+2\xi_n]) \approx 0.95. \]
- Also note that we could use \( \xi_n(0.95) \) in instead of "2" in if the number of samples \( n \) is small, cf., Student's t distribution.
- So, using the stopping criterion \( \xi_n \leq 0.1 \), the claim is then made that: the sample mean \( M_n \) is accurate to within 20% \([1-0.2, 1+0.2]\) of the true mean with 95% probability (or "19 times out of 20").
- Also, for a relative error of \( \xi_n = 0.1 \), the number of required samples is (by definition of \( \xi_n \)):
  \[ n = V_n/ (|M_n| \xi_n)^2 \approx 100 \ (\sigma/\mu)^2. \]
- Note how the number of samples increases with increasing variance \( \sigma^2 \) and decreasing absolute mean \( |\mu| \).
Stopping criterion for statistical confidence – important point in the confidence interval

- On the other hand, a stopping criterion could be based on whether an important point is present in the confidence interval
\[ [M_n - 2U_n^{0.5}/n^{0.5}, M_n + 2U_n^{0.5}/n^{0.5}] \]

- For example, if we are polling for a two-candidate election and are interested in whether one candidate will win the election by getting more than 50% of the vote:
  - for all \( i \), \( X_i \in \{0,1\} \) indicates whether the \( i^{th} \) sample is a vote for a certain candidate,
  - so we want to generate (poll for) enough samples \( n \) until 0.50 is not in the confidence interval;
  - i.e., if the confidence interval is \( > 0.5 \) (i.e., \( M_n - 2U_n^{0.5}/n^{0.5} > 0.5 \)), then we can say that the poll predicts that that candidate will win the election by getting a fraction \( M_n > 0.5 \) of the vote, with an error of \( \pm 2U_n^{0.5}/n^{0.5} \) “nineteen times out of twenty” (95% confidence).

Stopping criterion for statistical confidence (cont)

- If the election is a “dead heat” (EX=0.5), then we may find that for large all polling samples \( n \), 0.5 remains in the confidence interval.

- Finally, it might be necessary to produce a minimum number of samples \( n \) so that, e.g.,
  - the sample variance \( U_n > 0 \)
  - which may happen when the samples are discretely distributed and (by chance) the first \( n \) samples generated all have the same value.

- For example, in the two-candidate election just discussed, it may happen that the first 10 samples \( X \) are all zero (or all one).
Student’s t distribution

• Again, when the number of available samples $n$ is small, the following is used to define the confidence interval for small sample sizes $n$:

$$P(|(M_n - \mu)/(U_n^{0.5}/n^{0.5})| \leq \zeta_n(0.95)) \approx 0.95,$$

where the function $\zeta_n$ is defined by the Student’s t distribution with $n-1$ degrees of freedom (DoF) with PDF

$$f(x) = \left(1 + x^2/(n-1))^{-n/2}\Gamma(n/2)/[\Gamma((n-1)/2)(\pi(n-1))^{0.5}]\right.$$

• To see why (exercise):
  – **Assume** that the samples $X$ are i.i.d. N($\mu, \sigma^2$), so that $M_n$ is normally distributed too.
  – Prove that $M_n$ and $U_n$ are independent.
  – Note that $(n-1)U_n/\sigma^2$ has chi-square distribution with $n-1$ DoF.
  – Conclude that the t-statistic $(M_n - \mu)/(U_n^{0.5}/n^{0.5})$ is Student’s t distributed with $n-1$ DoF.
  – Now show as $n \to \infty$, Student’s t with n DoF converges to Gaussian.

Summary

• In summary, i.i.d. samples $X_i$ are produced from which a sequence of sample means $M_n$ and sample variances $V_n$ (or better, the unbiased $U_n$) are computed.

• The simulation may be terminated when the computed relative error $\xi_n$ reaches a predefined threshold or whether an important point is a member of the confidence interval.

• The accuracy of the resulting estimate $M_n$ of the quantity of interest $\mu = \text{EX}_i$ can be interpreted in terms of the confidence interval (via the Gaussian or Student’s t distributions).

• In practice, a certain minimum number of samples $n$ are required, e.g., to ensure $U_n > 0$. 

(c) 2013 George Kesidis
Statistical confidence: Example

- A great deal of current research activity in networking involves comparisons of the performance of competing mechanisms (devices, algorithms, protocols, etc.) by simulation or through prototypical deployment on the Internet.
- Suppose that n trials are conducted for each of two mechanisms in order to compare their performance, leading to a dataset \( \{ D_{k,i} \}_{i=1}^n \) for mechanism \( k \in \{ 1, 2 \} \).
- Also suppose that, for all i, the i-th trial or test (yielding data \( D_{1,i} \) and \( D_{2,i} \)) was conducted under arguably equal environmental conditions.
- Let \( X_i \equiv a_i(D_{1,i} - D_{2,i}) \) be the difference in the performance of the two mechanisms for the common environmental conditions of trial i, where scalars \( a_i > 0 \) may \( \neq 1 \) to make the outcomes of different trials comparable.
- That is, for each trial, an "apple-to-apple" comparison is made using coupled or paired observations.
- To assess whether \( \text{EX} > 0 \) (respectively, \( \text{EX} < 0 \) or \( \text{EX} = 0 \)), we simply compute the confidence interval accordingly to determine whether it is >0 (respectively, <0 or contains 0).

Deciding between two alternative claims

- Suppose that a series of measurements \( X_i \) are drawn from the Internet.
- A problem is that the Internet may be in different "states" of operation leading to samples \( X_i \) that are independent but not identically distributed.
- As a great simplification, suppose that the network can be in only one of two states indexed by \( j \in \{ 1, 2 \} \); also suppose the following.
- A batch of \( n \) samples \( X_i, 1 \leq i \leq n \), is taken while the network is in a single state but that state is not known.
- It is known that the probability that the network is in state 1 is \( p_1 \).
- Finally, \( \mu_j = E(X_i \mid \text{network state } j) \) for known values \( \mu_j, j \in \{ 1, 2 \} \).
- Without loss of generality, take \( \mu_1 < \mu_2 \).
- We wish to infer from the sample data \( X_i \) the state of the network.
- More specifically, we wish to minimize the probability of error \( P_e \) in our decision.
Deciding by using the CLT

- Note by the CLT, given that the network is in state \( j \), the sample mean \( M_n := n^{-1} \sum_{k=1}^{n} X_k \) is approximately normally distributed with mean \( \mu_j \) and variance \( \sigma_j^2 \).
- An unbiased, consistent estimate of this variance is \( U_n = (n-1)^{-1} \sum_{k=1}^{n} (X_k - M_n)^2 \).
- To determine the probability of decision error, we will condition on the state of the network:
  \[ P_e = P(\text{error} | \text{network state 1})p_1 + P(\text{error} | \text{network state 2})p_2 \]
  where \( p_2 = 1 - p_1 \).
- Consider a decision based on the comparison of the sample mean \( M_n \) with a threshold \( \theta \), where \( \mu_1 \leq \theta \leq \mu_2 \),
- so that the network is deemed to be in state 2 (having the higher mean) if \( M_n > \theta \); otherwise, the network is deemed to be in state 1.

Deciding by using the CLT (cont)

- A more concrete expression for the error probability ensues because we can approximate
  \[ P(\text{error} | \text{network in state 2}) = P(M_n \leq \theta | \text{network in state 2}) = \Phi((\mu_2 - \theta)/\sigma_2) \]
  where \( \Phi \) is 1 minus the CDF of a standard Gaussian distribution.
- Using a similar argument when conditioning on the network being in state 1, we arrive at the following expression:
  \[ P_e = \Phi((\mu_2 - \theta)/\sigma_2)p_2 + \Phi((\theta - \mu_1)/\sigma_1)p_1. \]
- Thus, the optimal value of \( \theta \) minimizes \( P_e \).
- This approach can be easily generalized to make decisions
  - where the two types of errors are weighted differently in \( P_e \) (i.e., not using \( p_i \)), or
  - among more than two alternative (mutually exclusive) hypotheses.
- Wald’s classical framework can decide between two alternative claims based on sequential independent observations.
Example: Attack/intrusion detection

- Consider a link carrying packets to a Web server.
- Suppose that, under normal operating conditions, the link will subject the Web server to a data rate of 4 Mbps.
- However, when the Web server is under a DDoS attack, the link will carry an average of 6 Mbps to the server.
- An IDS samples the link's data rate and determines whether the server is under attack.
- Assume known standard deviations in the data rate of 1 Mbps under normal conditions and of 1.5 Mbps under attack conditions.
- Finally, assume attack conditions exist 25% of the time.
- Find the value of the optimal decision threshold $\theta$ (compared against the sample mean) that minimizes the probability of decision error.

Example: Attack/intrusion detection (cont)

- Instead of minimizing the probability of decision error, suppose the probability that it was decided that the network is under attack when, in fact, it was not can be no more than 0.10, i.e.,
  \[ P_{FP}(\theta) = P(\text{decision error} | \text{no attack}) \leq 0.10. \]
- Again, this decision is based on the sample mean.
- Note that this event is called a false positive or type II error.
- Such false positives can be caused by legitimate "flash crowds" of temporary but excessive demand.
- Subject to this bound we wish to minimize the probability of missed detection ("type I" error) of an actual attack,
  \[ P_{MD}(\theta) = P(\text{decision error} | \text{attack}) \]
- That is, find \( \arg \min_\theta \{ P_{MD}(\theta) | P_{FP}(\theta) \leq 0.10 \} \)
- This is called a Neyman-Pearson test (of the sample mean with the computed value of $\theta$).
Chi-square test –
the “distance” between distributions

- Suppose N independent and identically distributed data samples x are
generated over a finite (or quantized finite) domain R.
- Their empirical PMF is d(r) = N⁻¹ ∑₁⁻¹ᵐₓ¹ N {x(i)=r} for r∈R.
- Suppose we want to check whether the empirical distribution is close
to a hypothetical one, say PMF p, that is the putative underlying
probability law generating the data samples.
- Pearson’s chi-square statistic
  \[ \sum_{r \in R} [d(r) - p(r)]^2 / p(r) \]
can be shown to asymptotically converge to a chi-square distribution.
- Exercise: explain how this result can be used to obtain a measure of
  confidence in the statement that d≈p.
- Other measures of the difference between a pair of distributions
  include the Kullback-Leibler (see Sanov’s theorem in Large Deviations)
  and the Kolmogorov-Smirnov.

Linear regression/MMSE-estimators revisited

- Suppose we have N independent and identically distributed data-
samples from an unknown distribution, where the nᵗʰ sample is
  \( (x_n, y_{1,n}, y_{2,n}, \ldots, y_{k,n}) \)
- We want the MMSE estimator \( \hat{Y} = a + b \) of X given \( Y = [Y_1, Y_2, \ldots, Y_k]^T \).
- One approach is to simply compute the unbiased sample means and
covariances as:
  \( M_{Y_i} = N⁻¹ \sum_{n=1}^{N} y_{i,n} \), \( M_X = N⁻¹ \sum_{n=1}^{N} x_n \),
  \( U_{Y_i,Y_j} = (N-1)⁻¹ \sum_{n=1}^{N} (y_{i,n} - M_Y)(y_{j,n} - M_Y) \) and
  \( U_{Y_i,X} = (N-1)⁻¹ \sum_{n=1}^{N} (y_{i,n} - M_Y)(x_n - M_X) \)
- If the random variables \( Y_1, Y_2, \ldots, Y_k \) are not linearly dependent, one
can show that the sample covariance matrix \( U_Y \) is a.s. nonsingular.
- Thus, we can compute linear MMSE estimator parameters \( a \) and \( b \) as
  \( a = U_Y^{-1} U_{Y,X} \) and \( b = M_X - U_Y^{-1} U_{Y,X} \)
Linear MMSE estimators revisited (cont)

- Specializing to the case $k=1$, suppose we are interested in testing whether $X$ and $Y=Y_1$ are uncorrelated, i.e., whether $E((X-EX)(Y-EY))=0$.
- Using the data, we want to determine whether $U_{YY}$ is sufficiently close to zero to confidently conclude that $X$ and $Y$ are uncorrelated.
- We can proceed as above and find the MMSE estimator $aY+b$ of $X$ given $Y$ as having parameters
  \[ a = U_YY^{-1} U_{YX} \quad \text{and} \quad b = M_Y - M_Y^T U_{CY} U_{YY}^{-1} U_{YX} ; \]
  so equivalently we want to determine whether the parameter $a$ is sufficiently close to zero to confidently conclude that $X$ and $Y$ are uncorrelated.
- Note that since parameters $a$ and $b$ are defined in terms of random samples, they themselves are random variables.

Linear MMSE estimators revisited (cont)

- Assuming the data is jointly Gaussian distributed, note that we can write $X = aY+b+R$ where $R$ is the Gaussian-distributed “residual” estimation error.
- Define the residual/error data of the linear MMSE estimator (or just the unqualified MMSE estimator in the Gaussian case) as
  \[ r_n = x_n - ay_n - b \]
- An unbiased estimate for the variance of $R$ is
  \[ \sigma_R^2 = (N-2)^{-1} \sum_{n=1}^{N} r_k^2 , \]
  where division by $N-2$ instead of $N$ accounts for the reduction in degrees of freedom due to the two parameters $a$ and $b$.
- An unbiased estimate of the variance of parameter $a$ is
  \[ \sigma_a^2 = \sigma_a^2 / \sum_{n=1}^{N} (y_n - M_Y)^2 \]
- So, the $t$-statistic (with $N-2$ DoF) for testing the hypothesis that $a=0$ is $(a-0)/\sigma_a$
**Remark:** Linear MMSE estimators

- For multidimensional linear regression (k>1), Hotelling’s $T^2$-test is a generalization of Student’s $t$-test.
- For linear prediction of a dependent time series, algorithms such as Levinson-Durbin or the discrete-time Kalman filter may be applied.

**Computer Exercise on Statistical Confidence**

- **Part I:** Generate $N\text{ i.i.d. Bernoulli}\{0,1\}$ samples $\{X_1, X_2, ..., X_n\}$ with $p=P(X=1)=0.51$ for one dataset (A) and $p=0.505$ for the other dataset (B); note again that the mean of such Bernoulli random variables is $EX = 0\cdot(1-p) + 1\cdot p = p$.
- For each set, estimate the mean (i.e., pretend you don’t already know $p=EX$) and compute the width of the 95% confidence interval for different values of $N$.
- Note that using the iterations above for empirical mean and variance, you don’t need to store the data as you generate it.
- Suppose our objective is to determine whether the mean is $>$ or $<$ 0.50, rather than a fixed percentage of the empirical mean.
- What value of $N$ is required for a 95% confidence interval whose width is .02 (i.e., ± .01) for the A dataset?
- What value of $N$ is required for a 95% confidence interval whose width is .01 (i.e., ± .005) for the B dataset?
- Argue the importance of this question if the datasets represent (explain how) the outcome of a close two-candidate election poll.
Computer Exercise on Statistical Confidence (cont)

- Generate the Bernoulli samples and update the estimates of the confidence intervals (in the same loop!), including the recursively updating the sample mean and standard deviation.
- Stop when (the entire confidence interval (CI) is either >0.5 or <0.5) and (the number of iterations is say > 20).
- **Part II:** Repeatedly generate independent sets of N i.i.d. Bernoulli{0,1} samples \( \{X_1, X_2, \ldots, X_N\} \), and for each set compute
  \[ L_N = \frac{(S_N - NEX)}{(N \text{ var}(X))^{0.5}} \]
- Plot histograms of 50 different (independent) samples \( M_N \) to visually verify that they are increasingly Gaussian in shape as \( N \) increases from 10 to 50 to 100 to 1000.
- See how the graphs change if the sample variance of \( X \) is used instead of the true variance in \( L_N \).
Outline

- Deterministic signals and linear, time-invariant (LTI) systems in continuous time
- The (zero state) impulse response and convolution
- The Fourier transform
- The transfer function of a LTI system
- Strong/strict stationarity of stochastic processes
- The autocorrelation function and wide-sense stationarity
- The power spectrum of a WSS process
- Broadband white noise
- Filtration of WSS processes
- Weiner-Hopf equations in discrete-time

Deterministic Dirichlet signals

- Consider a continuous-time, real-valued signal
  \( \{x(t) \mid t \in \mathbb{R}\} \), i.e., \( x : \mathbb{R} \rightarrow \mathbb{R} \).
- We assume all such signals \( x \) satisfy the Dirichlet conditions, allowing exchange of integral with limit, or switching the order of integration (Fubini’s theorem):
  - \( x \) is absolutely integrable, i.e., \( \int_{-\infty}^{\infty} |x(t)| \, dt < \infty \)
  - \( x \) has only finitely many discontinuities in any finite time-interval
  - \( x \) is bounded with only finitely many extrema in any finite time-interval (\( \sin(1/t) \) does not satisfy this in an interval containing 0).
- An even signal \( x \) is one such that
  \( x(t) = x(-t) \) for all \( t \in \mathbb{R} \)
- An odd signal \( x \) is one such that
  \( x(t) = -x(-t) \) for all \( t \in \mathbb{R} \)
- Any signal \( x \) may be decomposed into a sum of an even and odd signal:
  \( x(t) = 0.5[x(t)+x(-t)] + 0.5[x(t)-x(-t)] \), for all \( t \in \mathbb{R} \)
Deterministic periodic signals

• Let $\Delta_T$ be the operator that delays a signal by $T$ seconds (equivalently, advances a signal by $-T$ seconds), i.e.,
  $$(\Delta_T x)(t) = x(t-T), \text{ for all } t \in \mathbb{R}.$$  
• A signal $x$ is said to be periodic if there is a fixed time $T>0$ such that $x(t)=x(t-T)$ for all $t \in \mathbb{R}$, i.e., $x \equiv \Delta_T x$.
• The smallest such $T>0$ is known as the period of $x$.
• If $T$ is the period of $x$, then $1/T$ [Hz = 1/seconds] is the fundamental frequency of $x$, and any integer multiple $k/T$ is a harmonic of $x$.
• A constant signal $x$ has fundamental frequency 0 (called DC for “direct” current).

Sinusoidal signals

• A most important family of periodic signals are the sinusoids, i.e., of the form
  $$x(t)=A \cos(\omega t+\phi),$$
where the scalars:
  - $A>0$ is the amplitude,
  - $\omega$ is the frequency in radians/second ($2\pi/\omega$ is the period), and
  - $\phi$ is the phase.
• Note that we have referenced phase $\phi$ to cosine.
• Also note that $A>0$ without loss of generality, because for $A<0$ we can write $x(t) = -A \cos(\omega t+[\phi+\pi])$.
• The continuous-time complex-valued sinusoid is
  $$z(t) = A \exp(j[\omega t+\phi])=(Ae^{j\phi})e^{j\omega t}$$
where $j = \sqrt{-1} = (-1)^{0.5}$.
• Note that the real part of $z$ is $x$ and $Ae^{j\phi}$ is the phasor of $x$. 
Step, pulse and impulse signals

• The unit step function is
  \[ u(t) = 1_{\{t \geq 0\}} := \begin{cases} 1 & \text{if } t \geq 0 \\ 0 & \text{else} \end{cases} \]

• A rectangular pulse of height \( b \) and duty cycle (support) from \( r \) to \( s \) is
  \[ b[\Delta_u - \Delta_u] \]
  recall that
  \[ b[\Delta_u(t) - \Delta_u(t)] = b[u(t-r) - u(t-s)] = b \text{ if } s > t \geq r \text{ and } 0 \text{ else}. \]

• For positive \( a \ll 1 \), define the rectangular pulse
  \[ \delta_a = \frac{[u - \Delta_u]}{a}, \]
  where we've taken \( r = 0 \), \( s = a \), and \( b = 1/a \), i.e.,
  with support \( [0,a) \), height \( 1/a \), and hence area \( 1 \) for all \( a \).

• The impulse (Dirac delta function) \( \delta := \lim_{a \downarrow 0} \delta_a \)

• So by the above difference quotient, \( \delta = D \delta \), where \( D \) is the time-derivative operator.

Properties of the impulse

• Obviously by definition, \( \delta(t) = 0 \) for all \( t \neq 0 \).
• Thus, for any function \( x \), the function \( x\delta \) is such that
  \( (x\delta)(t) = x(t)\delta(t) = x(0)\delta(t), t \in \mathbb{R} \).
• Also by definition, \( \int_{-\infty}^{\infty} \delta(t-s)dt = 1 \) for all \( s \).
• Sampling property: For all functions \( x \) continuous at \( s \),
  \( \int_{-\infty}^{\infty} x(t)\delta(t-s)dt = x(s) \)
• For the special case of constant \( x = a \), we have \( \int_{-\infty}^{\infty} a\delta(t)dt = a \), where \( a\delta \) is notation for the impulse integrating to \( a \).
• We also see that \( \int_{-\infty}^{s} \delta(t)dt = u(s), s \in \mathbb{R} \).
• So, again, \( \delta = D u \).
• That is, at points of jump discontinuity, the derivative of a function will be an impulse weighted by the height of the jump.
Linear, time-invariant (LTI) systems

- Consider a single-input, single-output (SISO) system with input signal \( x \) and output signal \( y \).
- In this following, we define our signals on \( \mathbb{R} \) and assume zero initial conditions.
- That is, \( y \) is the zero (initial) state response (ZSR) of the system to \( x \).
- The system \( \mathcal{S} \) is said to be time-invariant if the response (ZSR) to \( \Delta_T x \) is \( \Delta_T y \) for all \( T \in \mathbb{R} \) and all (Dirichlet) input signals \( x \), where again \( y \) is the ZSR to \( x \), i.e., \( \mathcal{S} \) and \( \Delta_T \) commute: \( \forall T, \mathcal{S} \Delta_T = \Delta_T \mathcal{S} \).
- A linear combination of signals \( x_1 \) and \( x_2 \) is \( a_1x_1 + a_2x_2 \) where \( a_1, a_2 \in \mathbb{R} \) are scalars/ constants, i.e., \( (a_1x_1 + a_2x_2)(t) = a_1x_1(t) + a_2x_2(t), t \in \mathbb{R} \).
- The system \( \mathcal{S} \) is said to be linear if the response to \( a_1x_1 + a_2x_2 \) is \( a_1y_1 + a_2y_2 \) for all \( a_1, a_2 \in \mathbb{R} \) and all signals \( x_1, x_2 \), where \( y_k \) is the response to \( x_k \), i.e., \( \mathcal{S} \) commutes with signal amplification (homogeneity) and distributes over signal addition (superposition).

The ZSR of a LTI system is input convolved with impulse response

- Consider a specific LTI system and suppose its (ZSR) response to \( \delta_a \) (\( 0 < a \ll 1 \)) is \( h_a \), i.e., the system output, assuming zero initial conditions, is \( h_a \) when the system input is \( \delta_a \).
- So, as \( a \to 0 \), we get that the (zero state) impulse response is \( h = \lim_{a \to 0} h_a \).
- Now consider an arbitrary signal \( x \) and approximate it as a (piecewise constant) Riemann sum: \( x \approx x_a \) where
  \[
  x_a(t) = \sum_{k=-\infty}^{\infty} x(ka) \delta_a(t-ka) a
  \]
- Note that \( \delta_a(t-ka), t \in \mathbb{R} \), is a rectangular pulse of height 1 and support \( t \in [ka, (k+1)a) \).
- By linearity and time-invariance, the ZSR to \( x_a \) is
  \[
  y_a(t) = \sum_{k=-\infty}^{\infty} x(ka) h_a(t-ka) a
  \]
- As \( a \to 0 \), we see that the Riemann sum \( x_a \to x \) by the sampling property.
- So, the ZSR to \( x \) is \( \lim_{a \to 0} y_a = y \) where
  \[
  y(t) = \int_{-\infty}^{\infty} x(s) h(t-s) \, ds =: (x*h)(t),
  \]
  the convolution of \( x \) and \( h \).
The Fourier transform

- The Fourier transform of the continuous-time signal $x$ is the signal $X = \mathcal{F}\{x\}$ on frequencies $w \in \mathbb{R}$ (i.e., positive and negative frequencies),
  $$X(w) = \int_{-\infty}^{\infty} x(t) e^{-j\omega t} \, dt = \mathcal{F}\{x\}(w)$$
- The inverse Fourier transform is
  $$x(t) = (2\pi)^{-1} \int_{-\infty}^{\infty} X(w) e^{j\omega t} \, dw = \mathcal{F}^{-1}\{X\}(t)$$
- For example, if for some constant $T$, $x(t) = \delta(t-T)$, $t \in \mathbb{R}$, then by the sampling property, $X(w) = \exp(-jwT)$, $w \in \mathbb{R}$.
- Also, if $X(w) = 2\pi\delta(w-c)$, $w \in \mathbb{R}$, for some fixed frequency $c$, then $x(t) = \exp(jct)$, $t \in \mathbb{R}$.
- That is, impulse and complex exponential (sinusoid) are Fourier transform pairs.
- Exercise: Try to directly evaluate the Fourier-transform integral $x(t) = \exp(jct)$.

Properties of the Fourier transform

- The Fourier transform is a linear operator (because integration is linear), i.e., for all scalars $a, b \in \mathbb{R}$ and Dirichlet signals $x, y$,
  $$\mathcal{F}\{ax+by\} = a\mathcal{F}\{x\} + b\mathcal{F}\{y\}.$$  
- If the signal $x$ is real-valued (again with $X := \mathcal{F}\{x\}$), then the modulus (magnitude) of $X$, $|X(w)|$ is an even function of $w$ and the phase (angle) of $X$ is an odd function of $w$.
- The following time-derivative property of the Fourier transform immediately follows from integration-by-parts:
  If $X = \mathcal{F}\{x\}$ then $\mathcal{F}\{Dx\}(w) = jwX(w)$, $w \in \mathbb{R}$.
- The following time-shift property is proved by change-of-variable:
  $$\mathcal{F}\{\Delta_T x\}(w) = X(w)\exp(-jwT)$, $w \in \mathbb{R}$, for any fixed time $T \in \mathbb{R}$.
- Because the inverse Fourier transform is very similar in form to the Fourier transform, these properties have similar dual properties, e.g., the frequency-shift property:
  $$\mathcal{F}^{-1}\{\Delta w X\}(t) = x(t)\exp(jct)$, $t \in \mathbb{R}$.
An LTI system’s transfer function - the frequency domain perspective

• Again suppose \( y \) is the ZSR of a LTI system to \( x \), where the impulse response of the system is \( h \).

• Let \( X = \mathcal{F}\{x\} \) and \( H = \mathcal{F}\{h\} \).

• The Fourier transform of the ZSR \( y \) at frequency \( w \) is

\[
Y(w) = \int_{-\infty}^{\infty} y(t) e^{-jwt} \, dt
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(s) h(t-s) ds e^{-jwt} \, dt \quad \text{...since } y = x \ast h
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(s) e^{jws} h(t-s) e^{-jws} \, ds \, dt
\]

\[
= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x(s) e^{jws} h(t') e^{-jws} \, ds \, dt' \quad \text{where } t' = t-s
\]

• Thus, \( Y(w) = X(w)H(w) \), \( w \in \mathbb{R} \).

• So, convolution in the time domain corresponds to multiplication in the frequency domain.

• \( H \) is called the transfer function of the system, i.e., it transfers the input \( X \) to the output \( Y \) through multiplication.

Example: \( P(D)x = Q(D)y \)

• By iterative use of the time-derivative property and the linearity property of the Fourier transform, we can show that for any polynomial \( P \),

\[
\mathcal{F}\{P(D)x\}(w) = P(jw)X(w), \ w \in \mathbb{R}.
\]

• One can write the relationship between the input (current or voltage) signal \( x \) and (zero initial state) output signal \( y \) of a linear circuit as

\[
P(D)x = Q(D)y
\]

where here \( P \) and \( Q \) are polynomials.

• So, in the frequency domain, we can relate the input \( X \) to the (zero initial-state response) \( Y \) as

\[
P(jw)X(w) = Q(jw)Y(w), \ w \in \mathbb{R}.
\]
Example: $P(D)x = Q(D)y$ (cont)

- So, we can solve for $Y(w)/X(w)$ to get that the transfer function $H(w) = P(jw)/Q(jw)$.
- For example, the nodal equations (KCL in terms of node voltages) of the circuit below give
  
  $i + (x-y)/R + (0-y)/R = 0$ and $x-y = L \frac{di}{dt} = LDi$

  $\Rightarrow (x-y)/L + (Dx-Dy)/R - Dy/R = 0$

  $\Rightarrow P(D)x = 0.5Dx + (0.5R/L)x = Dy + (0.5R/L)y = Q(D)y$

- Thus, $H(w) = \frac{(0.5jw + 0.5R/L)}{(jw + 0.5R/L)}$ which can also be derived by simple voltage-division in the frequency domain, $H(w) = \frac{R}{R + (R || jwL)}$

Example: $P(D)x = Q(D)y$ (cont)

- So, if the input $X(w) = 2\pi \delta(w-c)$ for some fixed frequency $c$, i.e., $x(t) = \exp(jct)$, then

  $Y(w) = H(w) \ 2\pi \delta(w-c)$

  $= H(c) \ 2\pi \delta(w-c), \ w \in \mathbb{R}$.

  where we note that $H(c)$ is a constant function of frequency, $w$.

- Therefore, $y(t) = H(c) \exp(jct), t \in \mathbb{R}$.

- This is consistent with phasor analysis of linear and time-invariant systems.

- Since the ZSR to $\exp(jct)$ is $H(c) \exp(jct)$, the sinusoid $\exp(jct)$ is an eigenfunction of the LTI system with associated eigenvalue $H(c)$, and the ZSR is called an eigenresponse.
Strong/Strict Stationarity

• Now consider a continuous-time stochastic process $x$ over $\mathbb{R}$, i.e., $\{x(t) \mid t \in \mathbb{R}\}$.
• $x$ is said to be strictly stationary if all of its finite-dimensional distributions (FDDs) are time-shift invariant, i.e., if the joint CDF of $x(t(1)), \ldots, x(t(n))$ satisfies
  $$F_{t(1),t(2),\ldots,t(n)}(x) \equiv F_{t(1)+\tau,t(2)+\tau,\ldots,t(n)+\tau}(x)$$
  for all $\tau \in \mathbb{R}$, integers $n \geq 1$, $t(1), \ldots, t(n) \in \mathbb{R}$.
• If, for a certain integer $n$, the above statement is true only for $n$-dimensional distributions, then $x$ is said to be $n^{th}$ order stationary.
• Exercise: Argue that if $x$ is $n^{th}$ order stationary for some $n > 1$ then it is also $k^{th}$ order stationary for all $k \in \{1, 2, \ldots, n-1\}$.

Wide-sense stationarity

• A weaker definition for stationarity is wide-sense stationarity (WSS), which only requires:
  – the finite mean $E[x(t)] := m_X$ is a constant function of $t \in \mathbb{R}$; and
  – the autocorrelation $E[x(t)x(s)]$ depends on $t$ and $s$ only through their difference $t - s$, i.e.,
    $$r_X(s,t) := E[x(s)x(t)] = E[x(s + \tau)x(t + \tau)] \forall s, t, \tau \in \mathbb{R},$$
    where $r_X$ is the autocorrelation function of $x$, and
  – the second moment of the process at $t$ is finite, $r_X(t, t) < \infty \forall t \in \mathbb{R}$.
• That is, if $x$ is WSS,
  $$r_X(s,t) := r_X(s + \tau, t + \tau) \forall s, t, \tau \in \mathbb{R},$$
  and we can define its autocorrelation as a function of one variable,
  $$R_X(t - s) := r_X(0, t - s) = r_X(s, t) \forall s, t \in \mathbb{R},$$
  with $R_X(0) < \infty$.
• Exercise: Argue that any process that is $2^{nd}$ order strictly stationary is also wide-sense stationary.
Wide-sense stationarity (cont)

- Thus, the autocorrelation function of a WSS process \( X \) is well defined as
  \[
  R_X(\tau) := E[x(s)x(s + \tau)] \ orall s,\tau\in\mathbb{R},
  \]
- If \( x \) is complex valued, we take
  \[
  R_X(\tau) := E[x(s)x^*(s + \tau)] = (R_X(-\tau))^* \ orall s,\tau\in\mathbb{R},
  \]
  which gives
  \[
  R_X(0) := E|x(s)|^2 \ orall s,\tau\in\mathbb{R}.
  \]

Example of a WSS process

- Suppose \( \phi \sim \text{Uniform}[0,2\pi] \) where \( A,w > 0 \) are finite constants.
- Consider the stochastic process \( x(t) = A\cos(wt + \phi) \).
- So, for all \( t\in\mathbb{R} \):
  \[
  E[x(t)] = \int_0^{2\pi} A\cos(wt+z)/(2\pi) \ dz = A\sin(wt+z)/(2\pi)|_0^{2\pi} = A[\sin(wt+2\pi) - \sin(wt)]/(2\pi) = 0 = m_X
  \]
  Note how we used the uniform density \( f_\phi(z) = [u(z) - u(z-2\pi)]/(2\pi) \).
- Now, for all \( s\neq t\in\mathbb{R} \):
  \[
  E[x(t)x(s)] = \int_0^{2\pi} A\cos(wt+z)A\cos(ws+z)/(2\pi) \ dz = [A^2/(4\pi)] \int_0^{2\pi} [\cos(w(t+s)+2z) + \cos(w(t-s))] \ dz = [A^2/2] \cos(w(t-s)) = R_X(t-s)
  \]
  Similarly, we can check that \( R_X(0) = A^2/2 < \infty \).
- So, \( x \) is a WSS process.
Example of a WSS process that is not strictly stationary

- Suppose $x(t) = A\cos(\omega t + \phi)$, $t \in \mathbb{R}$, where $A, \omega$ are constants and $

\phi \sim \text{Uniform}(0, \pi/2, \pi, 3\pi/2)$, i.e., a discretely distributed 4-value phase (as in QPSK).
- Exercise: Check that $x(t)$ is a WSS process with $m_x$ and $R_X(\tau)$ as in the previous example.
- But $x(0) \in \{0, \pm 1\}$ a.s., whereas $x(\pi/(4\omega)) \in \{\pm 0.5\}$ a.s.
- So the (one-dimensional) distribution of $x(t)$ does depend on $t$, and the process is not stationary.
- This example from: M. Pursley. Random Processes in Linear Systems. Prentice-Hall.

Basic Properties of a WSS process, $x$

- If $x$ is real valued, $R_X$ is even, i.e., $R_X(t) = R_X(-t)$ $\forall t \in \mathbb{R}$, since $r_X(s,t) = r_X(t,s)$ $\forall s,t \in \mathbb{R}$
- $R_X(0) \geq |R_X(t)|$ $\forall t \in \mathbb{R}$, by the Cauchy-Schwarz inequality.
- Note: Some similar properties hold for the autocovariance of WSS process $X$, $c_X(s,t) := E[(x(s)-Ex(s))(x(t)-Ex(t))] = R_X(t-s) - m_X^2 =: C_X(t-s)$. 

(c) 2013 George Kesidis
Ergodicity – measuring m and R

- An ergodic process is one in which time averages converge to ensemble averages.
- For an ergodic WSS process \( x \),
  \[
  m_x = \mathbb{E}(x(t)) \quad \text{(ensemble average over sample paths at time } t) \\
  = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(x) \, dx \quad \text{(time average of one sample path)}
  \]
- Again, given one sample path \( x \) of an ergodic, Dirichlet, WSS process,
  \[
  R_x(t) = \lim_{T \to \infty} \frac{1}{2T} \int_{-T}^{T} x(s)x(s+t) \, ds, \quad \text{for all } t \in \mathbb{R}
  \]
- For a finite time-record of data (interval of observation) of \( x(t) \),
  \(-T < t < T < \infty\), we can approximate using the following unbiased and consistent estimators,
  \[
  m \approx (2T)^{1} \int_{-T}^{T} x(s) \, ds \\
  R_X(t) \approx (2T-t)^{-1} \int_{-T+t}^{T} x(s)x(s+t) \, ds
  \]

Power Spectral Density (PSD)

- Assume the autocorrelation \( R_x \) of a WSS, ergodic, Dirichlet signal \( x \) also satisfies the Dirichlet conditions, in particular \( \int_{-\infty}^{\infty} |R_x(t)| \, dt < \infty \).
- So, the Fourier transform \( \mathcal{F}\{R_X\} \) exists.
- Note that \( \forall t, \mathbb{E}^2(t) = \mathbb{E}(t) = R_x(0) = \mathcal{F}^{-1}\{R_X\}(0) = \int_{-\infty}^{\infty} S_X(w) \, dw/(2\pi) \),
  so that by ergodicity we arrive at Parseval’s theorem:
  \[
  \int_{-\infty}^{\infty} S_X(w) \, dw/(2\pi) = \lim_{T \to \infty} T^{1} \int_{-T/2}^{T/2} x^2(t) \, dt.
  \]
- Recall that if \( x \) is a voltage across, or current through, a resistor, then the instantaneous power dissipated by the resistor at time \( t \) is proportional to \( x^2(t) \).
- So, the right-hand-side of the previous display is proportional to the mean power dissipated by the resistor; the right-hand-side is also simply called the mean power of the signal \( x \).
- This explains why \( S_X \) is called a power spectral density, noting that \( [dw/(2\pi)] = \text{Hertz (cycles/s)} \) so that we can take \( [S_X] = \text{Watts/Hz} \).
LTI system response to a WSS input

• Suppose a WSS x is input to a LTI with (real-valued, deterministic) impulse response h, giving a ZSR y = x*h.
• Recall that the transfer function of the system is $H = \mathcal{F}\{h\}$ and we assume both x and h satisfy the Dirichlet conditions.
• **Theorem:** $y = x*h$ is WSS and $S_Y = |H|^2S_X$.
• **Proof:** Let $h(t) \equiv h(-t)$ and note that $\mathcal{F}\{h\}(w) = H(-w)$. ∀s,t∈ℝ:

$$r_Y(s,t) = Ey(s)y(t) = E \int_{-\infty}^{\infty} x(v)h(s-v)dv \int_{-\infty}^{\infty} x(u)h(t-u)du$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} Ex(v)x(u) h(s-v)h(t-u) du dv \quad \text{(Fubini)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(v-u) h(s-v)dv \ h(t-u)du \quad \text{(since X is WSS)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X(v') h(s-u-v')dv' \ h(t-u)du \quad \text{(sub v'=v-u)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X h(s-u)h(t-u)du$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} R_X h(u')h(t-s+u')du' \quad \text{(sub u'=s-u)}$$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (R_X h)*h(u')h(t-s+u')du'$$

$$= ((R_X h)*h)(s-t) = R_Y(t-s) = R_Y(s-t)$$

Thus, $y = x*h$ is WSS.

Now, by taking Fourier transform of $R_Y$, we get

$$S_Y(w) = \mathcal{F}\{R_X h\}(w)H(-w) = S_X(w)H(w)H(-w).$$

Since h is real valued, the modulus of $H$, $|H|$, is an even function and the angle of $H$ is odd.

Thus, $H(-w)$ is the conjugate of $H(w)$, and so $H(w)H(-w) = |H(w)|^2$.

Q.E.D.

(c) 2013 George Kesidis

319

(c) 2013 George Kesidis

320
Broadband White Noise

• Broadband white noise is a continuous-time WSS process \( W \) with:
  \[ \forall t \in \mathbb{R}, \]
  \[ m_W = \mathbb{E} w(t) = 0 \quad \text{(zero mean)} \]
  \[ R_W(t) = (N_0/2) \delta(t) \quad \text{for some constant } N_0 > 0. \]
• Thus, broadband white noise has constant power spectral density,
  \[ S_W(v) = N_0/2, \quad \text{for all frequencies } v \in \mathbb{R}. \]
• Though this idealized model of white noise has infinite mean power, it is nevertheless an extremely useful concept.

Power in a frequency band

• Note that we can compute the fraction of a signal \( x \)'s mean power that resides in a frequency range \([w_0 , w_1]\), where \( w_1 > w_0 > 0 \), as
  \[ \int_{w_0}^{w_1} S_X(w) \, dw / (2\pi) \]
• Exercise: show that the power of ideal white noise \( W \) in the frequency band range \([w_0 , w_1]\) (again where \( w_1 > w_0 > 0 \)) is
  \[ N_0 \, (w_1 - w_0). \]
• Note: The reason we divide the noise power by 2 in the definition of white noise is so that, by integrating over positive and negative frequencies, we don’t have to carry a factor of 2 in the previous expression.
Example: filtering white noise

- Find the PSD of the ZSR to broadband white noise of a LTI system with impulse response \( h(t)=\exp(-at), t \in \mathbb{R} \), for some constant \( a>0 \).
- First, the transfer function of the system is
  \[
  H(w)=\mathcal{F}\{h\}(w)=\frac{2a}{a^2+w^2}.
  \]
- So, the PSD of the WSS ZSR is
  \[
  |H(w)|^2\frac{N_0}{2}=\frac{2N_0a^2}{(a^2+w^2)^2}.
  \]
- Note how the power spectrum of the response is greatly attenuated for higher-magnitude frequencies, while this is not the case for lower-magnitude frequencies.
- That is, this system is a crude low-pass filter.
- **Exercise:** Find the PSD of the ZSR of this system to a sinusoid with additive white noise, \( x(t)=W(t)+A\cos(ct) \), for constants \( c,A>0 \). Hint: use linearity, the above result, and the eigenresponse. Repeat for the first WSS signal example with uniformly random phase.

Discrete-time signals and systems

- Discrete-time signals and LTI systems on \( \mathbb{Z} \)
- Kronecker delta (unit pulse) function and (zero state) system unit-pulse response
- LTI FIR filter in discrete time
- Discrete-time WSS processes
- Wiener-Hopf equations for minimum-MSE FIR estimator
Discrete-time signals

- A discrete-time signal is one whose domain is countable (1-1 mapping to integers \( \mathbb{Z} \) or a subset).
- We herein consider discrete-time signals \( x \) on \( \mathbb{Z} \):
  \[
x: \mathbb{Z} \rightarrow \mathbb{R}, \text{ i.e., } \forall k \in \mathbb{Z}, \ x[k] \in \mathbb{R}
\]
- For example,
  - the Kronecker delta (unit pulse): If \( k=0 \) then \( \delta[k]=1 \) else \( \delta[k]=0 \).
  - the exponential functions, \( \forall k \in \mathbb{Z}, x[k]=A r^k \)
- Note that any for signal \( x \) and time \( k \in \mathbb{Z} \),
  \[
x[k] = \sum_{m \in \mathbb{Z}} x[m] \delta[k-m] = \sum_{m=-\infty}^{\infty} x[m] \delta[k-m]
  = (x*\delta)[k], \text{ i.e., } \delta \text{ is the identity of convolution } *
\]
- Discrete and continuous convolution have similar properties.

Discrete-time Linear Time-Invariant Systems

- The properties of linearity and time-invariance for discrete-time systems are as for continuous-time systems.
- Again, for signal \( x \) and time \( k \in \mathbb{Z} \),
  \[
x[k] = \sum_{m \in \mathbb{Z}} x[m] \delta[k-m] = \sum_{m=-\infty}^{\infty} x[m] \delta[k-m]
\]
- Let \( h \) be the (zero state) unit-pulse response of a given linear-time invariant system.
- Thus, by linearity and time invariance, the ZSR to \( x \) is
  \[
y[k] = \sum_{m \in \mathbb{Z}} x[m] h[k-m] = (x*h)[k], \ k \in \mathbb{Z}
\]
- A causal finite impulse (unit pulse) response (FIR) filter is a system for which there is a time (lag) \( K \) such that
  \[
h[m] = 0 \text{ when } m<0 \text{ or } m\geq K, \text{ so that the ZSR to } x \text{ is}
  \]
  \[
y[k] = \sum_{m=k-K+1}^{k} x[m] h[k-m]
  = x[k]h[0]+x[k-1]h[1]...+x[k-K+1]h[K-1], \text{ for all time } k \in \mathbb{Z}
\]
Canonical realization of a FIR filter

- K-1 unit delays $z^{-1}$ for a K-parameter filter.
- Here $K=3$:

\[ x[k] \rightarrow z^{-1} x[k-1] \rightarrow h[1] \rightarrow z^{-1} x[k-2] \rightarrow h[2] \rightarrow \text{+} \rightarrow y[k] \]

WSS signals in discrete-time

- The notion of a WSS discrete time signal is as in continuous time.
- In particular, the auto-correlation function of a real-valued WSS signal $x$ is
  \[ R_x[k] = E x[m] x[m+k], \quad \forall m,k \in \mathbb{Z}. \]
- The cross-correlation function of two real-valued jointly WSS signals $x$ and $y$ is
  \[ R_{xy}[k] = E x[m] y[m+k], \quad \forall m,k \in \mathbb{Z}. \]
An estimation problem

- Suppose that we are given a signal $x$ and asked to design a FIR system $h$ of a certain order $K$ so that its output (ZSR) $y$ is a good approximation to a signal $v$, i.e., tracks $v$.
- Further suppose that $R_x$ and $R_{xv}$ are both known.
- Define the vector $h = (h[0] \ h[1] \ ... \ h[K-1])^T \in \mathbb{R}^K$.
- Define the vector ($\mathbb{R}^K$)-valued process:
  $x[k] = (x[k] \ x[k-1] \ ... \ x[k-K+1])^T$ for all time $k \in \mathbb{R}^K$
- Clearly, then $y[k] = h \cdot x[k]$, where we’ve used the dot/inner product.
- One can “design” $h$ by attempting to minimize the mean square error (MSE),
  $\varepsilon(h) := \mathbb{E}(v[k] - y[k])^2 = \mathbb{E}(v[k] - h \cdot x[k])^2$
- Note that under the WSS assumption, $\varepsilon$ does not depend on $k$.

Wiener-Hopf equations

- The (sufficient) first-order optimality conditions for this quadratic MSE function $\varepsilon(h)$ are:
  $0 = \partial \varepsilon(h)/\partial h[k] = \mathbb{E}((v[m] - h \cdot x[m])x[m-k])$, for $0 \leq k < K$.
- Thus,
  $0 = R_{xv}[k] \cdot \sum_{n=0}^{K-1} h[n] \mathbb{E}[x[m-n]x[m-k]]$, for $0 \leq k < K$
  $\Rightarrow R_{xv}[k] = \sum_{n=0}^{K-1} h[n] R_x[n-k]$, for $0 \leq k < K$
- These are the Wiener-Hopf equations.
- Define the $K$-vector $R_{sv} = (R_{sv}[0] \ R_{sv}[1] \ ... \ R_{sv}[K-1])$
- Define the $K \times K$ matrix $R_x$ whose entry in the $k$th row and $n$th column is $R_x[n-k]$.
- Note that since $x$ is assumed real-valued, $R_x$ is symmetric because
  $\forall m \in \mathbb{Z}, R_x[m] = R_x[-m]$.
- We can write the Wiener-Hopf equations more compactly as
  $R_{sv} = R_x h \Rightarrow h = R_x^{-1} R_{sv}$.
- So, if $R_x$ is non-singular, we can find the MMSE FIR filter $h$ to track $v$. 