Information Theory and Networks

Lecture 3: Revision: Probability Theory

Matthew Roughan
<matthew.roughan@adelaide.edu.au>
http://www.maths.adelaide.edu.au/matthew.roughan/Lecture_notes/InformationTheory/

School of Mathematical Sciences,
University of Adelaide

October 9, 2013

Part I

A Recap of Probability

> It is impossible for a Die, with such determin'd force and direction, not to fall on such determin'd side, only I don't know the force and direction which makes it fall on such determin'd side, and therefore I call it Chance, which is nothing but the want of art....
>
> *John Arbuthnot*
> *(in the preface of 'Of the Laws of Chance', 1692)*

Probability

Topics you should be familiar with:

- Axiomatic Probability
- Random Variables
- Distributions: focussing on discrete distributions
- Conditional Probability
- Expectations
- Jensen's Inequality
- Markov Chains (but we will cover these later)

Section 1

Probability Axioms

Section 1
Probability Axioms

Probability Axioms

- What does "we get heads with probability half" mean?
  - it could mean that we believe a coin flipped a number of times n will come up heads $n/2$ times **but that patently isn't true**
  - it could mean that in the long run it comes up heads half the time **but what if we know an event will only occur once?**
  - what about a more fundamental approach
- Axioms state things that we believe are intuitively true, but not provable.
  - they are the starting points of reasoning
- Probability axioms are defined on sets
  - I assume you know set notation and rules
  - we talk about subsets of elements as events
  - we will denote the certain event Ω
  - we will talk about the probability of event $E \subseteq \Omega$ being $P(E)$.

Probability Axioms
- What does "we get heads with probability half" mean?
  - it could mean that we believe a coin flipped a number of times n will come up heads $n/2$ times **but that patently isn't true**
  - it could mean that in the long run it comes up heads half the time **but what if we know an event will only occur once?**
  - what about a more fundamental approach
- Axioms state things that we believe are intuitively true, but not provable.
  - they are the starting points of reasoning
- Probability axioms are defined on sets
  - I assume you know set notation and rules
  - we talk about subsets of elements as events
  - we will denote the certain event Ω
  - we will talk about the probability of event $E \subseteq \Omega$ being $P(E)$.

Probability Axioms

The axioms are

1. $P(E) \in \mathbb{R}^+$, i.e., $P(E)$ is real, and non-negative
2. $P(\Omega) = 1$, i.e., probability of the entire sample space is 1.
3. Any countable, sequence of disjoint events $E_1, E_2, \ldots$ satisfies

$$P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$$

Probability Axioms
The axioms are
1. $P(E) \in \mathbb{R}^+$, i.e., $P(E)$ is real, and non-negative
2. $P(\Omega) = 1$, i.e., probability of the entire sample space is 1.
3. Any countable, sequence of disjoint events $E_1, E_2, \ldots$ satisfies
$$P(E_1 \cup E_2 \cup \cdots) = \sum_i P(E_i).$$

Immediate Consequences

1. Monotonicity
   $$\text{if } A \subseteq B \text{ then } P(A) \le P(B)$$
2. Empty set ϕ has probability zero: $P(\phi) = 0$.
3. Probabilities are all bounded: $0 \le P(E) \le 1$.
4. Complementary probabilities
   $$P(E^c) = P(\Omega \backslash E) = 1 - P(E).$$
5. Addition law
   $$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
6. Law of total probability: given a countable partition of Ω into $E_1, E_2, \ldots,$, we can write the probability
   $$P(A) = \sum_i P(A \cap E_i).$$

Immediate Consequences
1. Monotonicity
$$\text{if } A \subseteq B \text{ then } P(A) \le P(B)$$
2. Empty set ϕ has probability zero: $P(\phi) = 0$.
3. Probabilities are all bounded: $0 \le P(E) \le 1$.
4. Complementary probabilities
$$P(E^c) = P(\Omega \backslash E) = 1 - P(E).$$
5. Addition law
$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$
6. Law of total probability: given a countable partition of Ω into $E_1, E_2, \ldots,$, we can write the probability
$$P(A) = \sum_i P(A \cap E_i).$$

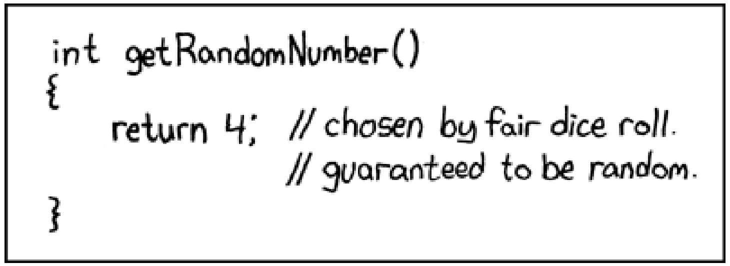
Section 2

Random Variables

Section 2

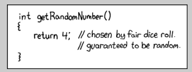
Random Variables

Random Variables



http://xkcd.com/221/

Random Variables



http://xkcd.com/221/

Random Variables

Intuitively a Random Variable (RV) is a variable, e.g., X, that takes a random numerical value.

1. probability is defined on sets, but a lot of the time we just want a random number
2. we'll use X, Y, and Z to mean a RV, and x, y, and z to mean the values they take.

But they still have to satisfy the axioms of probability, and we want a firm foundation to work on.

Random Variables
Intuitively a Random Variable (RV) is a variable, e.g., X, that takes a random numerical value.
1. probability is defined on sets, but a lot of the time we just want a random number
2. we'll use X, Y, and Z to mean a RV, and x, y, and z to mean the values they take.
But they still have to satisfy the axioms of probability, and we want a firm foundation to work on.

Random Variables: formal approach

Consider an experiment with a sample space Ω.

1. We take a set of subsets of this called $\sigma(\Omega)$
   1. technically this should be a σ-algebra, but we won't need to deal too much with this here
2. A RV is a mapping from $\sigma(\Omega)$ to the reals, e.g.

$$X : \sigma(\Omega) \to \mathbb{R}.$$

So for each possible outcome we might measure, we assign a number, which is our RV.

Random Variables: formal approach
Consider an experiment with a sample space Ω.
1. We take a set of subsets of this called $\sigma(\Omega)$
   1. technically this should be a σ-algebra, but we won't need to deal too much with this here
2. A RV is a mapping from $\sigma(\Omega)$ to the reals, e.g.
$$X : \sigma(\Omega) \to \mathbb{R}.$$
So for each possible outcome we might measure, we assign a number, which is our RV.

Cumulative Distribution Function (CDF)

Now we can assign probabilities to RV values.

1. Because they are on a number line we exploit the ordering, and use the CDF defined thus

$$F_X(x) = P(X \leq x) = P\big(\{e \in \Omega | X(e) \leq x\}\big).$$

2. Properties
   1. $F_X(-\infty) = 0$ and $F_X(\infty) = 1$
   2. Nondecreasing: $x_1 \leq x_2$ implies $F_X(x_1) \leq F_X(x_2)$
   3. Right continuous

$$\lim_{\epsilon \to 0} F_X(x + \epsilon) = F_X(x), \text{ for } \epsilon > 0$$

   but not necessarily left-continuous.

3. The density function (where defined) is the derivative of the CDF, but we have to be careful about this because it isn't always defined.
   1. in particular for discrete distributions its more useful to work with the probability mass function.

Section 3

Discrete Distributions

Probability Mass Function (PMF)

In this course, we will mostly deal with discrete distributions:

1. intuitively RV takes on a (countable) set of discrete values x_i;
2. CDF is piecewise constant;

In these cases, the PMF is sometimes more useful

$$p_X(x_i) = P(X = x_i) = F_X(x_i) - F_X(x_i^-),$$

where $x_i^- =$ the left-hand limit.

We often simplify when the context is clear, e.g.,

$$p_X(x_i) = p(x_i) = p_i.$$

Joint Distributions

Given two (discrete) random variables X and Y, we write the joint PMF

$$p_{X,Y}(x, y) = P(X = x \text{ and } Y = y).$$

Example Distributions

1. Uniform: $\Omega = \{1, 2, \ldots, n\}$

$$p(k) = 1/n.$$

2. Bernoulli: $\Omega = \{0, 1\}$

$$p(1) = p, \text{ and } p(0) = 1 - p = q$$

3. Binomial (sum of n independent Bernoulli trials): $\Omega = \{0, 1, \ldots, n\}$

$$p(k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

4. Poisson: $\Omega = \mathbb{Z}^+$, the non-negative integers:

$$p(k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Laplace's principle of insufficient reason

> The theory of chance consists in reducing all the events of the same kind to a certain number of cases equally possible, that is to say, to such as we may be equally undecided about in regard to their existence, and in determining the number of cases favorable to the event whose probability is sought. The ratio of this number to that of all the cases possible is the measure of this probability, which is thus simply a fraction whose numerator is the number of favorable cases and whose denominator is the number of all the cases possible.
>
> *Pierre Simon Laplace*

- If we don't know any better, then assume a uniform distribution.
  - this is used a lot, e.g., probability of the Ace of Hearts
  - its pretty fundamental, but also axiomatic in nature
- Called "Principle of Indifference" by John Maynard Keynes (1921)

Section 4

Conditional Probability

Section 4
Conditional Probability

Probability Axioms (revisited)

We left one thing out of the axioms: conditional probability

1. Define $P(A|B)$ as
   1. probability of A conditioned on B
   2. probability of A given B
   3. probability of A will occur, given that we know B has, or will occur
   4. probability of A accounting for the evidence B
2. Missing Axiom

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad \text{if } P(B) > 0.$$

Probability Axioms (revisited)
We left one thing out of the axioms: conditional probability
Define $P(A|B)$ as
probability of A conditioned on B
probability of A given B
probability of A will occur, given that we know B has, or will occur
probability of A accounting for the evidence B
Missing Axiom
$P(A|B) = \frac{P(A \cap B)}{P(B)}$, if $P(B) > 0$.

Independence

Two events A and B are said to be independent iff

$$P(A|B) = P(A),$$

Equivalently:

1. $P(B|A) = P(B)$
2. $P(A \cap B) = P(A)P(B)$

Independence
Two events A and B are said to be independent iff
$P(A|B) = P(A)$,
Equivalently:
1. $P(B|A) = P(B)$
2. $P(A \cap B) = P(A)P(B)$

Bayes' Law

$$P(A|B) = P(B|A)\frac{P(A)}{P(B)}.$$

- much more could be said about this
- interpretation of this rule has caused arguments amongst statisticians for centuries

Bayes' Law
$P(A|B) = P(B|A)\frac{P(A)}{P(B)}$
- much more could be said about this
- interpretation of this rule has caused arguments amongst statisticians for centuries

Law of Total Probability (reprise)

Given a countable partition of Ω into $E_1, E_2, \ldots,$, we can write the probability

$$
\begin{aligned}
P(A) &= \sum_i P(A \cap E_i) \\
&= \sum_i P(A|E_i) P(E_i).
\end{aligned}
$$

Probabilistic Chain Rule

$$
\begin{aligned}
P(A_n \cap A_{n-1} \cap \cdots \cap A_1) &= P(A_n | A_{n-1} \cap \cdots \cap A_1) P(A_{n-1} \cap \cdots \cap A_1) \\
&= P(A_n | A_{n-1} \cap \cdots \cap A_1) \\
&\quad \times P(A_{n-1} | A_{n-2} \cap \cdots \cap A_1) \\
&\quad \times P(A_{n-2} \cap \cdots \cap A_1)
\end{aligned}
$$

So

$$
P(A_3 \cap A_2 \cap A_1) = P(A_3 | A_2 \cap A_1) P(A_2 | A_1) P(A_1).
$$

Section 5

Expectations

Section 5
Expectations

Expectation

The expectation of a (discrete) random variable taking values x_i is defined to be

$$E[X] = \sum_i x_i p_X(x_i).$$

- the expectation is commonly called the average or mean
- Example: expectation of a uniform random variable U:

$$E[U] = \frac{1}{n}\sum_{i=1}^{n} u_i.$$

- Example: expectation of a Poisson random variable

$$E[X] = \sum_{k=0}^{\infty} kp(k) = \lambda e^{-\lambda} \sum_{k=0}^{\infty} k\frac{\lambda^{k-1}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

Expectation
The expectation of a (discrete) random variable taking values x_i is defined to be
$$E[X] = \sum_i x_i p_X(x_i).$$
- the expectation is commonly called the average or mean
- Example: expectation of a uniform random variable U:
$$E[U] = \frac{1}{n}\sum_{i=1}^{n} u_i.$$
- Example: expectation of a Poisson random variable
$$E[X] = \sum_{k=0}^{\infty} kp(k) = \lambda e^{-\lambda} \sum_{k=0}^{\infty} k\frac{\lambda^{k-1}}{k!} = \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda.$$

Expectations of functions

- We can take expectations of a function of a random variable

$$E[g(X)] = \sum_i g(x_i)p(x_i).$$

- Examples:
  - $$E[-\log_2(X)] = -\sum_i \log_2(x_i)p(x_i).$$
  - An indicator function is

$$I_A(X) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

The expectation of an indicator is

$$E[I_A(X)] = P(x \in A).$$

Expectations of functions

- We can take expectations of a function of a random variable

$$E[g(X)] = \sum_i g(x_i)p(x_i).$$

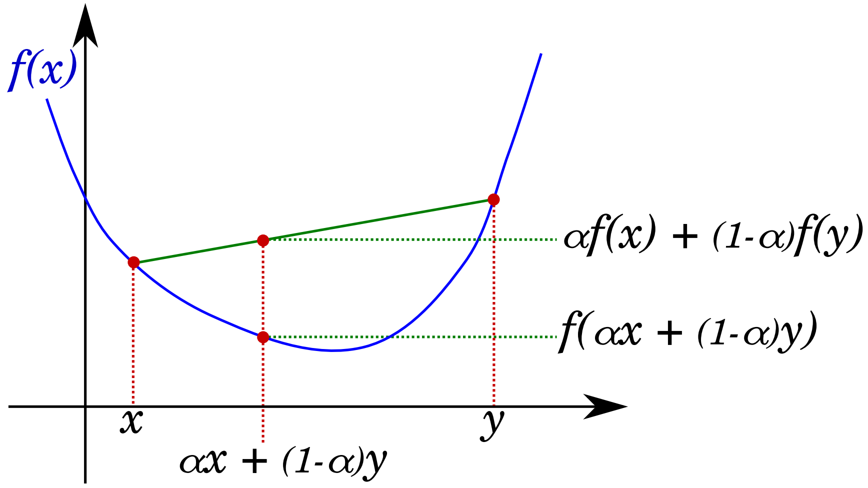
- One approach to defining higher-order moments of a distribution is to say the pth moment is

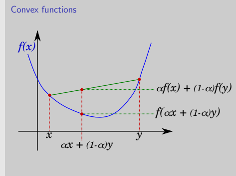
$$m_p = E[X^p] = \sum_i x_i^p p(x_i).$$

or the pth central moment is

$$\mu_p = E\left[(X - E[X])^p\right] = \sum_i (x_i - E[X])^p p(x_i).$$

Convex functions

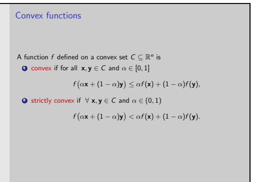




Convex functions

A function f defined on a convex set $C \subseteq \mathbb{R}^n$ is

1. convex if for all $\mathbf{x}, \mathbf{y} \in C$ and $\alpha \in [0,1]$
$$f\left(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\right) \leq \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}),$$
2. strictly convex if $\forall\ \mathbf{x}, \mathbf{y} \in C$ and $\alpha \in (0,1)$
$$f\left(\alpha \mathbf{x} + (1-\alpha)\mathbf{y}\right) < \alpha f(\mathbf{x}) + (1-\alpha) f(\mathbf{y}).$$



Jensen's Inequality

For any random variable X and convex function $g(\cdot)$ Jensen's inequality states:

$$g\left(E[X]\right) \leq E\left[g(X)\right]$$

and if $g(\cdot)$ is strictly convex, then equality only holds for X deterministic.
Examples

- $E[X^2] \geq E[X]^2$
  e.g., consider $X = \{-1, 1\}$ each with probability $1/2$
- $E\left[|X|\right] \geq \left|E[X]\right|$
  e.g., again consider $X = \{-1, 1\}$ each with probability $1/2$
- $E[-\log(X)] \geq -\log\left(E[X]\right)$

Gibbs' Inequality

Take two probability mass functions $p_i = p(x_i)$ and $q_i = q(x_i)$ defined over the same set of events x_i. Then

$$-\sum_i p_i \log_2 p_i \leq -\sum_i p_i \log_2 q_i,$$

with equality iff $p_i = q_i$.
Proof: use Jensen on the negative log of random variables taking values $y_i = q_i/p_i$ with probability p_i.

Properties of Expectation

1. Jensen: for convex $g(\cdot)$
$$g\big(E[X]\big) \le E\big[g(X)\big].$$
2. Indicators
$$E[I_A(X)] = P(x \in A).$$
3. Linearity
$$E[aX + bY] = aE[X] + bE[Y].$$

Conditional Expectation

We can also define the expectation conditional on an event, e.g.,
$$E[X|Y = y] = \sum_i x_i p(x_i|y).$$

Conditional expectations behave in most ways like normal expectations, just WRT to a different probability measure.

Conditional Expectation as a RV

If the event we are conditioning on is a RV itself, then the conditional expectation $E[X|Y]$ is a RV too:

- It is a function mapping the values of Y to real numbers
- We can talk about probabilities, expectations and so on
- If X and Y are independent

$$E[X|Y] = E[X]$$

- If X is completely determined by Y, e.g., $X = g(Y)$ then

$$E[X|Y] = E[g(Y)|Y] = g(Y) = X$$

- Also

$$E\big[E[X|Y]\big] = E[X]$$

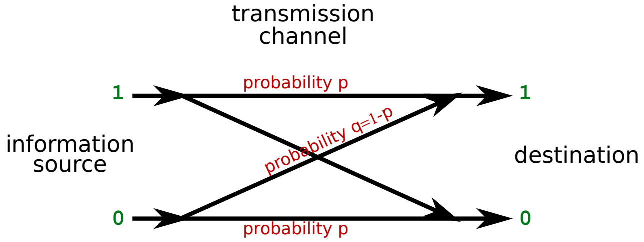
Conditional Expectation as a RV
If the event we are conditioning on is a RV itself, then the conditional expectation $E[X|Y]$ is a RV too:
- It is a function mapping the values of Y to real numbers
- We can talk about probabilities, expectations and so on
- If X and Y are independent
$E[X|Y] = E[X]$
- If X is completely determined by Y, e.g., $X = g(Y)$ then
$E[X|Y] = E[g(Y)|Y] = g(Y) = X$
- Also
$E\big[E[X|Y]\big] = E[X]$

Section 6

Examples

Section 6
Examples

Example 1: binary communications system

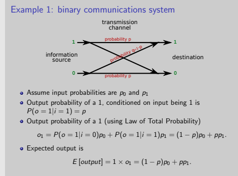


- Assume input probabilities are p_0 and p_1
- Output probability of a 1, conditioned on input being 1 is $P(o = 1|i = 1) = p$
- Output probability of a 1 (using Law of Total Probability)

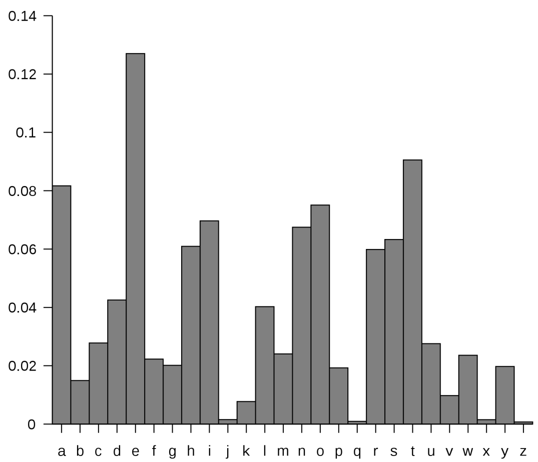
$$o_1 = P(o = 1|i = 0)p_0 + P(o = 1|i = 1)p_1 = (1 - p)p_0 + pp_1.$$

- Expected output is

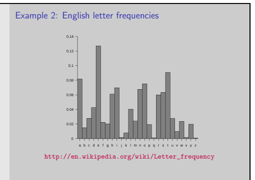
$$E\left[output\right] = 1 \times o_1 = (1 - p)p_0 + pp_1.$$



Example 2: English letter frequencies



http://en.wikipedia.org/wiki/Letter_frequency



a	8.167 %	b	1.492 %	c	2.782 %
d	4.253 %	e	12.702 %	f	2.228 %
g	2.015 %	h	6.094 %	i	6.966 %
j	0.153 %	k	0.772 %	l	4.025 %
m	2.406 %	n	6.749 %	o	7.507 %
p	1.929 %	q	0.095 %	r	5.987 %
s	6.327 %	t	9.056 %	u	2.758 %
v	0.978 %	w	2.360 %	x	0.150 %
y	1.974 %	z	0.074 %		

http://en.wikipedia.org/wiki/Letter_frequency
And the data and figure is directly available at
http://en.wikipedia.org/wiki/File:English_letter_frequency_(alphabetic).svg

Assignment

Learn Morse Code. You will need to be able to translate Morse Code into text, from memory (though not in real time) by next lecture.
There are some helpful web sites:

- http://www.learnmorsecode.com/
- http://www.wikihow.com/Learn-Morse-Code
- http://www.justlearnmorsecode.com/

As I said, don't worry about timing, we'll be writing out translations, but I will test you.
Write a short (less than 1/2 a page) description of how redundancy in language and the Morse code interact. For instance, what happens if a telegraph line is noisy, and how efficient is Morse code?

Assignment

Learn Morse Code. You will need to be able to translate Morse Code into text, from memory (though not in real time) by next lecture.
There are some helpful web sites:

- http://www.learnmorsecode.com/
- http://www.wikihow.com/Learn-Morse-Code
- http://www.justlearnmorsecode.com/

As I said, don't worry about timing, we'll be writing out translations, but I will test you.
Write a short (less than 1/2 a page) description of how redundancy in language and the Morse code interact. For instance, what happens if a telegraph line is noisy, and how efficient is Morse code?

Further reading I

- Rick Durrett, *Probability: Theory and examples*, 3rd ed., Thomson, 2005.
- William Feller, *An introduction to probability theory and its applications*, second ed., vol. I, John Wiley and Sons, New York, 1971.
- ______, *An introduction to probability theory and its applications*, second ed., vol. II, John Wiley and Sons, New York, 1971.
- Karlin, *A first course in stochastic processes*, Academic Press, 1969.
- J.F.C. Kingman and S.J. Taylor, *Introduction to measure and probability*, Cambridge University Press, 1966.
- Henry Stark and John Woods, *Probability and random processes with applications to signal processing*, 3rd ed., Prentice Hall, 2002.