

# 41901: Probability and Statistics

Nick Polson

Fall, 2019

October 1, 2019

# Getting Started

- ▶ Syllabus  
<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41900/>
- ▶ General Expectations
  1. Read the notes / Practice
  2. Be on schedule
  3. DeGroot and Schervish: Probability and Statistics
  4. Polson and Scott: AIQ

# Course Expectations

## Course Expectations :

Homework: 20% Assignments. Handed in at Class

I encourage you to do assignments in groups.

Otherwise it's no fun!

Grading is ✓ −, ✓ , ✓ +.

Final: 80% Week 11

Grading: PhD course

## Table of Contents

Chapter 1 : Probability, Bayes	Slide 5
Chapter 2: Distributions and Expectations	Slide 107
Chapter 3: Bayesian Statistics	Slide 152
Chapter 4: Modern Regression Methods	Slide 170
Chapter 5 : Hierarchical Models	Slide 197
Chapter 6 : Bayesian Finance	Slide 239
Chapter 7 : AI	Slide 273
Chapter 8 : Deep Learning	Slide 295

# Introduction

## 1. W1-W7 Probability (DeGroot and Schervish)

Chap 1: Probability

Chap 2: Bayes

Chap 3: Random Variables

Chap 4: Expectations

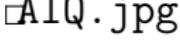
Chap 5: Special Distributions

Chap 6: Hierarchical Models

## 2. W8-W10 AI and Deep Learning

AIQ

## AIQ: People and Machines Smarter Together

- ▶ Florence Nightingale 
- ▶ Sir Isaac Newton
- ▶ Grace Hopper

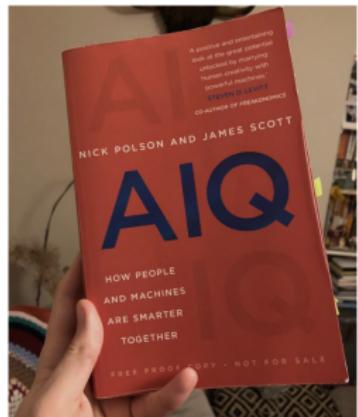
February 27, 2019 · Technology, Book Notes

## Book Notes: AIQ by Nick Polson and James Scott

f t h

Machine Learning and Artificial Intelligence are words that appear in public discourse on an almost daily basis as we discuss and debate their impressive, and sometimes not so impressive, achievements. However it's rare to find simple explanations about how such things were achieved so that they can seem "almost indiscernible from magic". Even as an engineer, I was unaware of some of the basic ideas behind the most common machine learning algorithms. *AIQ* does a fantastic job of demystifying these and many others with some thoughtful discussion about what the future might hold for us.

Here are some of my notes and thoughts on the book and some links to further reading. I was lucky enough to given an advance copy by a friend, [Sophie Christopher](#), from Penguin Random House - thanks Sophie!



# Outline

1. **Probability** (axioms, subjective, utility, laws, conditional, Bayes theorem, applications, cdf's and pdf's)
2. **Distributions** (discrete, continuous, binomial, poisson, gamma, weibull, beta, exponential family, wishart)  
Transformations and Expectations (distribution of functions, expectations, mgf's, convergence, conditional and marginal, bivariate normal, inequalities and identities)
3. **Modern Regression Methods** (Ridge, Lasso)
4. **Bayesian Methods** (Hierarchical Models, Shrinkage, Asset Allocation, Brownian Motion)
5. **AI and Deep Learning** (NNs, SGD, Dropout, Applications)

# Probability: 41901

## **Week 1: Probability, Risk and Utility**

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Overview

## Probability and Paradoxes

1. Birthday Problem
2. Exchange Paradox
3. Probability: Axioms and Subjectivity
4. Expected Utility: Preferences
5. Risk
6. Probability and Psychology
7. St. Petersburg Paradox
8. Allais Paradox
9. Kelly Criterion

Reading: DeGroot and Schervish: Chapter 1,1-47.

# Birthday Problem

## Example

*“.. for the ‘one chance in a million’ will undoubtedly occur, with no less and no more than its appropriate frequency, however surprised we may be that it should occur to us” (Fisher, 1937)*

1. Birthdays: 23 people for a 50/50 chance of a match.
2. Almost Birthdays:  
Only need 13 people for a Birthday match to within a day.
3. Multiple Matches: 88 people for a triple match

# Solution

Easier to calculate  $P(\text{no match})$

- ▶ The first person has a particular birthday

$$P(\text{no match}) = 1 \times \frac{364}{365} \times \dots \times \frac{364 - N + 1}{365}$$

Substituting  $N = 23$  you get  $P(\text{no match}) = \frac{1}{2}$ .

- ▶ General rule-of-thumb:  
 $N$  people,  $c$  the number of categories

Then  $N = 1.2\sqrt{c}$  for a 50/50 chance.

- ▶ To apply this to Birthdays we use:  $c = 365$  and  $N = 23$   
For near Birthdays  $c = 121$  and  $N = 1.2\sqrt{121} = 13$ .

- We can calculate

$$\begin{aligned} P(\text{no match}) &= \prod_{i=1}^{N-1} \left(1 - \frac{i}{c}\right) \\ &= \exp\left(\sum_{i=1}^{N-1} \log\left(1 - \frac{i}{c}\right)\right) \approx \exp\left(-\frac{N^2}{2c}\right) \end{aligned}$$

where  $\log\left(1 - \frac{i}{c}\right) \approx -\frac{i}{c}$  for  $i \ll c$  and  $\sum_{i=1}^{N-1} i = \frac{1}{2}N(N - 1)$ .

# Probability

Probability is a language designed to communicate uncertainty.

It's immensely useful, and there's only a few basic rules.

1. If an event  $A$  is certain to occur, it has probability 1, denoted  $P(A) = 1$
2. Either an event  $A$  occurs or it does not.

$$P(\text{not } A) = 1 - P(A)$$

3. If two events are mutually exclusive (both cannot occur simultaneously) then  $P(A \text{ or } B) = P(A) + P(B)$
4.  $P(A \text{ and } B) = P(A \text{ given } B)P(B) = P(A|B)P(B)$

# Odds

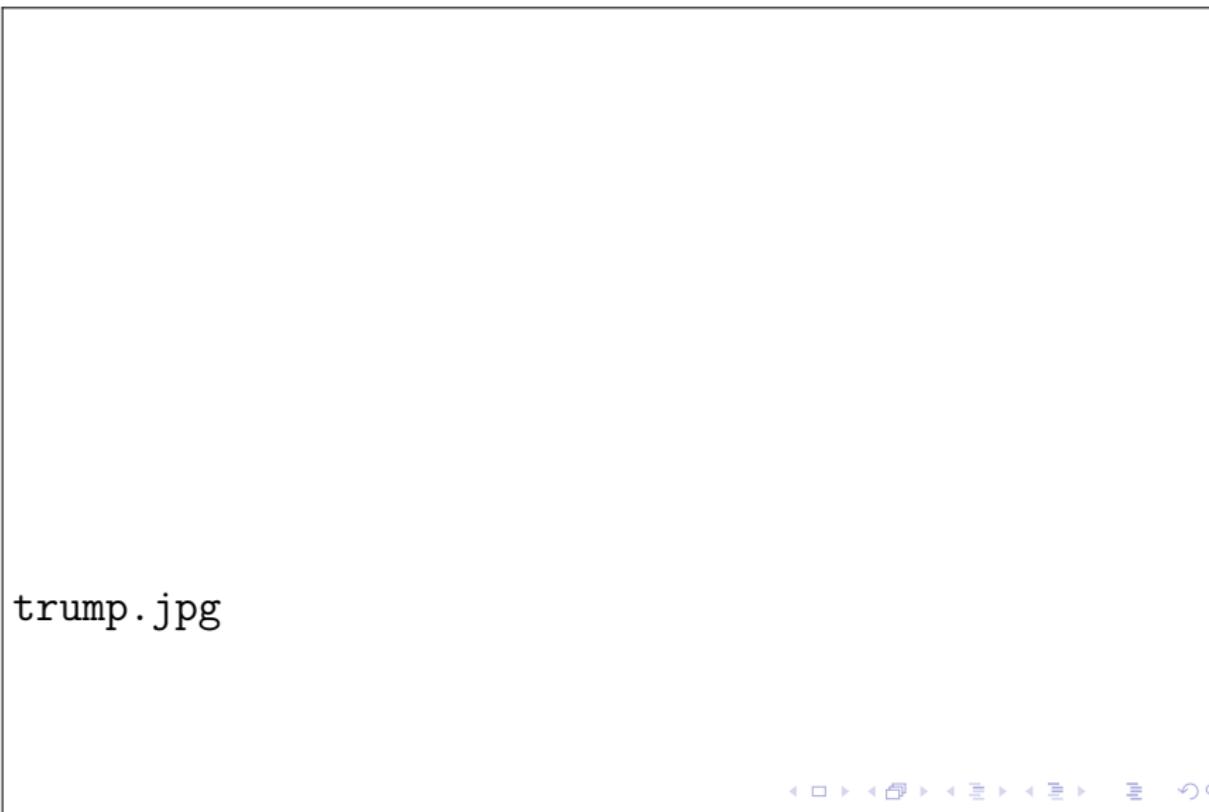
We can express probabilities in terms of Odds via

$$O(A) = \frac{1 - P(A)}{P(A)} \text{ or } P(A) = \frac{1}{1 + O(A)}$$

- ▶ For example if  $O(A) = 1$  then for ever \$1 bet you will payout \$1.  
An event with probability  $\frac{1}{2}$ .
- ▶ If  $O(A) = 2$  or  $2 : 1$ , then for a \$1 bet you'll payback \$3.  
In terms of probability  $P = \frac{1}{3}$ .

# US Presidential Election 2020

Oddschecker



trump.jpg

# Bayes Theorem

Many problems in decision making can be solved using Bayes rule.

- ▶ Rule-based decision making. Artificial Intelligence.
- ▶ It's counterintuitive! But gives the "right" answer.

Bayes Rule:

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

Law of Total Probability:

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

# Bayes Theorem

If you take the test and the result is positive, you are really interested  
in the question: Given that you tested positive, what is the chance  
you have the disease?

Let  $D = 1$  indicate you have a disease

Let  $T = 1$  indicate that you test positive for it

# Bayes Theorem

We have a probability table

		D	
		0	1
T	0	0.9702	0.001
	1	0.0098	0.019

Bayes Probability

$$\mathbb{P}(D = 1 | T = 1) = \frac{0.019}{(0.019 + 0.0098)} = 0.66$$

# Envelope Paradox

The following problem is known as the “exchange paradox”.

- ▶ A swami puts  $m$  dollars in one envelope and  $2m$  in another. He hands one envelope to you and one to your opponent.

The amounts are placed randomly and so there is a probability of  $\frac{1}{2}$  that you get either envelope.

You open your envelope and find  $x$  dollars. Let  $y$  be the amount in your opponent's envelope.

# Envelope Paradox

- ▶ You know that  $y = \frac{1}{2}x$  or  $y = 2x$ . You are thinking about whether you should switch your opened envelope for the unopened envelope of your friend. It is tempting to do an expected value calculation as follows

$$E(y) = \frac{1}{2} \cdot \frac{1}{2}x + \frac{1}{2} \cdot 2x = \frac{5}{4}x > x$$

Therefore, it looks as if you should switch no matter what value of  $x$  you see. A consequence of this, following the logic of backwards induction, that even if you didn't open your envelope that you would want to switch!

# Bayes Rule

- ▶ Where's the flaw in this argument? Use Bayes rule to update the probabilities of which envelope your opponent has! Assume  $p(m)$  of dollars to be placed in the envelope by the swami.
- ▶ Such an assumption then allows us to calculate an odds ratio

$$\frac{p(y = \frac{1}{2}x|x)}{p(y = 2x|x)}$$

concerning the likelihood of which envelope your opponent has.

- ▶ Then, the expected value is given by

$$E(y) = p\left(y = \frac{1}{2}x \mid x\right) \cdot \frac{1}{2}x + p(y = 2x|x) \cdot 2x$$

and the condition  $E(y) > x$  becomes a decision rule.

# Apple Watch Series 4 ECG and Bayes' Theorem

The Apple Watch Series 4 can perform a single-lead ECG and detect atrial fibrillation. The software can correctly identify 98% of cases of atrial fibrillation (true positives) and 99% of cases of non-atrial fibrillation (true negatives).

However, what is the probability of a person having atrial fibrillation when atrial fibrillation is identified by the Apple Watch Series 4?

Bayes' Theorem:

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

Predicted		atrial fibrillation	no atrial fibrillation
atrial fibrillation		1960	980
no atrial fibrillation		40	97020

$$0.6667 = \frac{0.98 \cdot 0.02}{0.0294}$$

The conditional probability of having atrial fibrillation when the Apple Watch Series 4 detects atrial fibrillation is about 67%.

# Probability

Subjective probability is very different from the (non-operational) approach based on long-run averages.

1. Kreps (1989) Notes on the Theory of Choice
2. Ramsey (1926) Truth and Probability
3. deFinetti (1930) Theory of Probability
4. Kolmogorov (1931) Probability
5. von Neumann and Morgenstern (1944) Theory of Economic Games and Behaviour
6. Savage (1956) Foundations of Statistics

Principle of Coherence:

*A set of subjective probability beliefs must avoid sure loss*

# Probability and Psychology

How do people form probabilities or expectations in reality?  
Psychologists have categorized many different biases that people have in their beliefs or judgments.

**Loss Aversion** The most important finding of Kahneman and Tversky is that people are loss averse.

Utilities are defined over gains and losses rather than over final (or terminal) wealth, an idea first proposed by Markowitz. This is a violation of the EU postulates. Let  $(x, y)$  denote a bet with gain  $x$  with probability  $y$ . To illustrate this subjects were asked

*In addition to whatever you own, you have been given \$1000, now choose between the gambles  $A = (1000, 0.5)$  and  $B = (500, 1)$ .*

$B$  was the more popular choice.

# Example

The same subjects were then asked: *In addition to whatever you own, you have been given \$2000, now choose between the gambles*  
 $C = (-1000, 0.5)$  and  $D = (-500, 1)$ .

- ▶ This time  $C$  was more popular.
- ▶ The key here is that their final wealth positions are identical yet people chose differently. The subjects are apparently focusing only on gains and losses.

When they are not given any information about prior winnings, they choose  $B$  over  $A$  and  $C$  over  $D$ . Clearly for a risk averse people this is the rational choice.

- ▶ This effect is known as loss aversion.

# Representativeness

**Representativeness** When people try to determine the probability that evidence  $A$  was generated by model  $B$ , they often use the representative heuristic. This means that they evaluate the probability by the degree to which  $A$  reflects the essential characteristics of  $B$ .

A common bias is *base rate neglect* or ignoring prior evidence. For example, in tossing a fair coin the sequence *HHTHTHHTHH* with seven heads is likely to appear and yet people draw conclusions from too few data points and think 7 heads is representative of the true process and conclude  $p = 0.7$ .

# Expected Utility (EU) Theory

## Normative

Let  $P, Q$  be two probability distributions or *risky gambles/lotteries*.  
 $pP + (1 - p)Q$  is the *compound* or *mixture* lottery.

The rational agent (You) will have preferences between gambles.

- ▶ We write  $P \succeq Q$  if and only if You strictly prefer  $P$  to  $Q$ . If two lotteries are *indifferent* we write  $P \sim Q$ .
- ▶ EU – a number of plausible axioms – completeness, transitivity, continuity and independence – then preferences are an expectation of a utility function.
- ▶ The theory is a *normative* one and not necessarily *descriptive*. It suggests how a rational agent should formulate beliefs and preferences and not how they actually behave.
- ▶ Expected utility  $U(P)$  of a risky gamble is then

$$P \succeq Q \iff U(P) \geq U(Q)$$

# Attitudes to Risk

The solution depends on your *risk* preferences:

- ▶ *Risk neutral*: a risk neutral person is indifferent about fair bets.  
Linear Utility
- ▶ *Risk averse*: a risk averse person prefers certainty over fair bets.

$$\mathbb{E}(U(X)) < U(\mathbb{E}(X)) .$$

Concave utility

- ▶ *Risk loving*: a risk loving person prefer fair bets over certainty.

Depends on **your preferences**.

# Ellsberg Paradox

*Probability is counter-intuitive!!!*

- ▶ Two urns
  1. 100 balls with 50 red and 50 blue.
  2. A mix of red and blue but you don't know the proportion.
- ▶ Which urn would you like to bet on?
- ▶ People don't like the “uncertainty” about the distribution of red/blue balls in the second urn.

# Likelihood of Death

120 Stanford graduates:

Heart disease		34
cancer		23
Other natural causes		35
Total	58	actual 92
accident		5
homicide		1
other unnatural causes		2
Total	32	actual 8

- ▶ The  $P$ 's don't even sum up to one!  
People vastly overestimate probability of violent death

# St. Petersburg Paradox

What are you willing to pay to enter the following game?

- ▶ I toss a fair game and when the first head appears, on the  $T$ th toss, I pay you  $\$2^T$  dollars.
- ▶ First, probability of first head on  $T$ th toss is  $2^{-T}$

$$\begin{aligned} E(X) &= \sum_{T=1}^{\infty} 2^T 2^{-T} \\ &= 2(1/2) + 4(1/4) + 8(1/8) + \dots \\ &= 1 + 1 + 1 + \dots \rightarrow \infty \end{aligned}$$

- ▶ Bernoulli (1754) constructed utility to value bets with  $E(u(X))$ .

# Allais Paradox

You have to make a choice between the following gambles

- ▶ First compare the “Gambles”

$\mathcal{G}_1$ : \$ 5 million with certainty

$\mathcal{G}_2$ : \$ 25 million  $p = 0.10$

\$ 5 million  $p = 0.89$

\$ 0 million  $p = 0.01$

- ▶ Now choose between the Gambles

$\mathcal{G}_3$ : \$ 5 million  $p = 0.11$

\$ 0 million  $p = 0.89$

$\mathcal{G}_4$ : \$ 25 million  $p = 0.10$

\$ 0 million  $p = 0.90$

Fact: If  $\mathcal{G}_1 \geq \mathcal{G}_2$  then  $\mathcal{G}_3 \geq \mathcal{G}_4$  and vice-versa.

# Solution: Expected Utility

Given (subjective) probabilities  $P = (p_1, p_2, p_3)$ . Write  $E(u|P)$  for expected utility.

- ▶ Without loss of generality we can set  $u(0) = 0$  and for the high prize set  $u(\$25 \text{ million}) = 1$ . Which leaves one free parameter  $u = u(\$5 \text{ million})$ .
- ▶ Hence to compare gambles with probabilities  $P$  and  $Q$  we look at the difference

$$E(u|P) - E(u|Q) = (p_2 - q_2)u + (p_3 - q_3)$$

- ▶ For comparing  $\mathcal{G}_1$  and  $\mathcal{G}_2$  we get

$$E(u|\mathcal{G}_1) - E(u|\mathcal{G}_2) = 0.11u - 0.1$$

$$E(u|\mathcal{G}_3) - E(u|\mathcal{G}_4) = 0.11u - 0.1$$

The order is the same, given *your*  $u$ .

- ▶ If your utility satisfies  $u < 0.1/0.11 = 0.909$  you take the “riskier” gamble.

# Utility Functions

## Power and log-utilities

- ▶ Constant relative risk aversion (CRRA).
- ▶ Advantage that the optimal rule is unaffected by wealth effects.  
The CRRA utility of wealth takes the form

$$U\gamma(W) = \frac{W^{1-\gamma} - 1}{1 - \gamma}$$

- ▶ The special case  $U(W) = \log(W)$  for  $\gamma = 1$ .  
This leads to a myopic Kelly criterion rule.

# Kelly Criterion

Kelly Criterion corresponds to betting under binary uncertainty.

- ▶ Consider a sequence of i.i.d. bets where

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

The optimal allocation is  $\omega^* = p - q = 2p - 1$ .

- ▶ Maximising the expected long-run growth rate leads to the solution

$$\max \omega \mathbb{E} (\ln(1 + \omega W_T)) = p \ln(1 + \omega) + (1 - p) \ln(1 - \omega)$$

$$\leq p \ln p + q \ln q + \ln 2 \text{ and } \omega^* = p - q$$

# Kelly Criterion

If one believes the event is certain i.e.  $p = 1$ , then one bets all wealth and *a priori* one is certain to double invested wealth.

If the bet is fair, i.e.  $p = \frac{1}{2}$ , one bets nothing,  $\omega^* = 0$ , due to risk-aversion.

- ▶ Let  $p$  denote the probability of a gain and  $O = (1 - p)/p$  the odds. We can generalize the rule to the case of asymmetric payouts  $(a, b)$  where

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

- ▶ Then the expected utility function is

$$p \ln(1 + b\omega) + (1 - p) \ln(1 - a\omega)$$

- ▶ The optimal solution is

$$\omega^* = \frac{bp - aq}{ab} = \frac{p - q}{\sigma}$$

# Kelly Criterion

- ▶ If  $a = b = 1$  this reduces to the pure Kelly criterion.
- ▶ A common case occurs when  $a = 1$ . We can now interpret  $b$  as the odds  $O$  that the market is willing to offer the invest if the event occurs and so we write  $b = O$ . The rule becomes

$$\omega^* = \frac{p \cdot O - q}{O}$$

## Example

- ▶ Two possible market opportunities: one where it offers you 4/1 when you have personal odds of 3/1 and a second one when it offers you 12/1 while you think the odds are 9/1. In expected return these two scenarios are identical both offering a 33% gain. In terms of maximizing long-run growth, however, they are not identical.

# Example

- ▶ Table 1 shows the Kelly criteria advises an allocation that is twice as much capital to the lower odds proposition: 1/16 weight versus 1/40.

Market	You	$p$	$\omega^*$
4/1	3/1	1/4	1/16
12/1	9/1	1/10	1/40

Table: Kelly rule

- ▶ The optimal allocation  $\omega^* = (pO - q)/O$  is

$$\frac{(1/4) \times 4 - (3/4)}{4} = \frac{1}{16} \text{ and } \frac{(1/10) \times 12 - (9/10)}{12} = \frac{1}{40}$$

# Assignment 1

## Probability and Paradox

1. Envelope Paradox
2. Galton's Paradox
3. Prosecutors' Fallacy
4. St Petersburg Paradox
5. Kelly Criterion
6. OJ Simpson Case

# Probability: 41901

## Week 2: Bayes Rule

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Topics

1. Conditional Probability
2. Bayes Rule
3. Prisoner's Dilemma/Monte Hall
4. Using Probability as Evidence
5. Island Problem and DNA Evidence
6. Combining Evidence
7. Prosecutors and base-rate fallacies.

Reading: DeGroot and Schervish: Chapter 2, p.49-97.

# Conditional Probability

## Bayes Rule

In its simplest form.

- ▶ Two events  $\mathcal{A}$  and  $\mathcal{B}$ . Bayes rule

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

- ▶ Law of Total Probability

$$P(B) = P(B|A)P(A) + P(B|\bar{A})P(\bar{A})$$

Hence we can calculate the denominator of Bayes rule.

# Prisoner's Dilemma

Three prisoners  $\mathcal{A}, \mathcal{B}, \mathcal{C}$ .

Each believe are equally likely to be set free.

Prisoner  $\mathcal{A}$  goes to the warden  $\mathcal{W}$  and asks if s/he is getting axed.

- ▶ The Warden can't tell  $\mathcal{A}$  anything about him.
- ▶ He provides the new information:  $\mathcal{WB} = "B$  is to be executed"

# Prisoner's Dilemma

Uniform Prior Probabilities:

Prior	$\mathcal{A}$	$\mathcal{B}$	$\mathcal{C}$
$\mathcal{P}(\text{Pardon})$	0.33	0.33	0.33

Posterior: Compute  $P(\mathcal{A}|\mathcal{WB})$ ?

What happens if  $\mathcal{C}$  overhears the conversation?

Compute  $P(\mathcal{C}|\mathcal{WB})$ ?

# The Game Show Problem

Named after the host of the long-running TV show, *Let's make a Deal*.

- ▶ A **contestant** is given the choice of 3 doors.  
There is a **prize (a car, say)** behind one of the doors and something worthless behind the other two doors: two goats.
- ▶ The optimal strategy is **counter-intuitive**

# Puzzle

The game is as follows:

- ▶ You pick a door.
- ▶ Monty then opens one of the other two doors, revealing a goat.
- ▶ You have the choice of switching doors.

Is it advantageous to switch?

Assume you pick door  $A$  at random. Then  $P(A) = (1/3)$ .

You need to figure out  $P(A|MB)$  after Monte reveals  $B$  is a goat.

21 Kevin Spacey: Monte Hall

# Probability as Evidence

**evidence:** known facts about criminal (e.g. blood type, DNA, ...)

**suspect:** matches a trait with evidence at scene of crime

Let  $\mathcal{G}$  denote the event that the suspect is the criminal.

- ▶ Bayes computes the conditional probability of guilt

$$P(\mathcal{G}|\text{evidence})$$

Evidence  $\mathcal{E}$ : suspect and criminal possess a common trait

# Probability as Evidence

Bayes Theorem yields

$$P(\mathcal{G}|\text{evidence}) = \frac{P(\text{evidence}|\mathcal{G})P(\mathcal{G})}{P(\text{evidence})}$$

In terms of **relative odds**

$$\frac{P(\mathcal{I}|\text{evidence})}{P(\mathcal{G}|\text{evidence})} = \frac{P(\text{evidence}|\mathcal{I})}{P(\text{evidence}|\mathcal{G})} \frac{P(\mathcal{I})}{P(\mathcal{G})}$$

# Bayes Factors

There are two terms:

1. Prior Odds of Guilt  $O(\mathcal{G}) = P(\mathcal{I})/P(\mathcal{G})$  ?

How many people on the island?

Sensitivity “what if” analysis?

2. The Bayes factor

$$\frac{P(\text{evidence}|\mathcal{I})}{P(\text{evidence}|\mathcal{G})}$$

is common to all observers and updates everyone's initials odds

# Prosecutor's Fallacy

The most common fallacy is confusing

$$P(\text{evidence}|\mathcal{G}) \text{ with } P(\mathcal{G}|\text{evidence})$$

- ▶ Bayes rule yields

$$P(\mathcal{G}|\text{evidence}) = \frac{P(\text{evidence}|\mathcal{G})p(\mathcal{G})}{P(\text{evidence})}$$

- ▶ Your assessment of  $P(\mathcal{G})$  will matter.

# Island Problem

Suppose there's a criminal on a island of  $N + 1$  people.

- ▶ Let  $I$  denote innocence and  $G$  guilt.
- ▶ Evidence  $E$ : the suspect matches a trait with the criminal.
- ▶ The probabilities are

$$p(E|I) = p \text{ and } p(E|G) = 1$$

# Bayes factor

Bayes factors are likelihood ratios

- ▶ The Bayes factor is given by

$$\frac{p(E|I)}{p(E|G)} = p$$

- ▶ If we start with a uniform prior distribution we have

$$p(I) = \frac{1}{N+1} \text{ and } odds(I) = N$$

- ▶ Priors will matter!

# Island Problem (contd)

Posterior Probability related to Odds

$$p(I|y) = \frac{1}{1 + odds(I|y)}$$

- ▶ *Prosecutors' fallacy*

The posterior probability  $p(I|y) \neq p(y|I) = p$ .

- ▶ Suppose that  $N = 10^3$  and  $p = 10^{-3}$ . Then

$$p(I|y) = \frac{1}{1 + 10^3 \cdot 10^{-3}} = \frac{1}{2}$$

The odds on innocence are  $odds(I|y) = 1$ .

There's a 50/50 chance that the criminal has been found.

# Sally Clark Case: Independence or Bayes?

Sally Clark was accused and convicted of killing her two children  
They could have both died of SIDS.

- ▶ The chance of a family which are non-smokers and over 25 having a SIDS death is around 1 in 8,500.
- ▶ The chance of a family which has already had a SIDS death having a second is around 1 in 100.
- ▶ The chance of a mother killing her two children is around 1 in 1,000,000.

# Bayes or Independence

## 1. Under Bayes

$$\begin{aligned} P(\text{both SIDS}) &= P(\text{first SIDS}) P(\text{Second SIDS|first SIDS}) \\ &= \frac{1}{8500} \cdot \frac{1}{100} = \frac{1}{850,000} \end{aligned}$$

The  $\frac{1}{100}$  comes from taking into account genetics.

## 2. Independence, as the court did, gets you

$$P(\text{both SIDS}) = (1/8500)(1/8500) = (1/73,000,000)$$

## 3. By Bayes rule

$$\frac{p(I|E)}{p(G|E)} = \frac{P(E \cap I)}{P(E \cap G)}$$

$P(E \cap I) = P(E|I)P(I)$  needs discussion of  $p(I)$ .

# Comparison

- ▶ Hence putting these two together gives the odds of guilt as

$$\frac{p(I|E)}{p(G|E)} = \frac{1/850,000}{1/1,000,000} = 1.15$$

In terms of posterior probabilities

$$p(G|E) = \frac{1}{1 + O(G|E)} = 0.465$$

- ▶ If you use independence

$$\frac{p(I|E)}{p(G|E)} = \frac{1}{73} \text{ and } p(G|E) \approx 0.99$$

The suspect looks guilty.

# OJ Simpson

The O.J. Simpson trial was possibly the trial of the century  
The murder of his wife Nicole Brown Simpson, and a friend, Ron Goldman, in June 1994 and the ensuing trial dominated the TV networks

- ▶ DNA evidence and probability:  $p(E|I)$
- ▶ Bayes Theorem:  $p(G|E)$
- ▶ Prosecutor's Fallacy:  $p(G|E) \neq p(E|G)$

Odds ratio with  $I = \bar{G}$  gives

$$\frac{p(I|E)}{p(G|E)} = \frac{p(E|I)}{p(E|G)} \frac{p(I)}{p(G)}$$

Prior odds conditioned on background information.

# OJ Simpson: Bayes Theorem

Suppose that you are a juror in a murder case of a husband who is accused of killing his wife.

The husband is known to have battered her in the past.

Consider the three events:

1.  $G$  “husband murders wife in a given year”
2.  $M$  “wife is murdered in a given year”
3.  $B$  “husband is known to batter his wife”

# OJ Simpson: Bayes Theorem

- ▶ Only 1/10th of one percent of husbands who batter their wife actually murder them.  
Conditional on eventually murdering their wife, there is a one in ten chance it happens in a given year.

In a given year, there are about 5 murders per 100,000 of population in the United States.

In 1994, 5000 women were murdered, 1500 by their husband  
Given a population of 100 million women at the time

$$p(M|I) = \frac{3500}{1 \times 10^8} \approx \frac{1}{30,000}.$$

We'll also need  $p(M|I, B)$

# OJ Simpson: Prosecutor's Fallacy

Let  $G$  = Guilt and  $E$  = Evidence

*Prosecutor's Fallacy:*  $P(G|E) \neq P(E|G)$ .

DNA evidence gives  $P(E|I)$  – the  $p$ -value.

What's the “match probability” for a rare event

Bayes theorem in Odds

$$\frac{p(G|M, B)}{p(I|M, B)} = \frac{p(M|G, B)}{p(M|I, B)} \frac{p(G|B)}{p(I|B)}$$

# OJ Simpson: Bayes Theorem

By assumption,

- ▶  $p(M|G, B) = 1$
- ▶  $p(M|I, B) = \frac{1}{20,000}$
- ▶  $p(G|B) = \frac{1}{1000}$  and so

$$\frac{p(G|B)}{p(I|B)} = \frac{1}{10,000}$$

Therefore,

$$\frac{p(G|M, B)}{p(I|M, B)} = 2 \text{ and } p(G|M, B) = \frac{2}{3}$$

More than a 50/50 chance that your spouse murdered you!

# Fallacy $p(G|B) \neq p(G|B,M)$

Alan Dershowitz stated to the press: in any given year

*“Fewer than 1 in 2000 of batterers go on to murder their wives”.*

- ▶ Now estimate  $p(M|\bar{G}, B) = p(M|\bar{G}) = \frac{1}{20,000}$ .
- ▶ The Bayes factor is then

$$\frac{p(G|M, B)}{p(\bar{G}|M, B)} = \frac{1/2000}{1/20,000} = 10$$

which implies posterior probabilities

$$p(\bar{G}|M, B) = \frac{1}{1+10} \text{ and } p(G|M, B) = \frac{10}{11}$$

Hence its over 90% chance that O.J. is guilty based on this information!

Dershowitz intended this information to exonerate O.J.

# Proving Innocence

DNA evidence is very good at proving innocence:

$$O(G|E) = \frac{p(y|I)}{p(y|G)} O(G)$$

- ▶ A small Bayes factor  $p(y|I)/p(y|G)$  implies  $O(G|E)$  is also small.
- ▶ Hence, we see that no match implies innocence.
- ▶ Match probability is the probability that unrelated individual chosen randomly from a reference population will match the profile.

# More than One Criminal

Let  $E$  two criminals and suspect matches  $C$ .

- ▶ Rare blood type ( $p = 0.1$ ) and Common blood type  $C$  ( $p = 0.6$ )
- ▶ Evidence for the prosecution?
- ▶ Bayes:  $P(C|I) = 0.6$  and  $P(C|G) = 0.5$

$$\mathcal{BF} = \frac{P(C|G)}{P(C|I)} = \frac{0.5}{0.6} = 0.93 < 1!$$

Evidence for the Defense. Posterior Odds on Guilt are decreased.

# Base Rate Fallacies

“Witness” 80 % certain saw a “checker” C taxi in the accident.

- ▶ What’s your  $P(C|E)$  ?
- ▶ Need  $P(C)$ . Say  $P(C) = 0.2$  and  $P(E) = 0.8$ .
- ▶ Then your posterior is

$$P(C|E) = \frac{0.8 \cdot 0.2}{0.8 \cdot 0.2 + 0.2 \cdot 0.8} = 0.5$$

Therefore  $O(C) = 1$  a 50/50 bet.

# Updating Fallacies

Most people don't update quickly enough in light of new data  
Wards Edwards 1960s

When you have a small sample size, Bayes rule still updates probabilities

- ▶ Two players: either 70 % A or 30 % A
- ▶ Observe  $A$  beats  $B$  3 times out of 4.
- ▶ What's  $P(A = 70\% \text{ player})$  ?

# Sequential Learning: Bellman

*Bellman's principle of optimality*

*An optimal policy has the property that whatever the initial state and initial decision are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decision. (Bellman, 1956)*

*In Math!!*

For a finite horizon in discrete time with no discounting is given by

$$V(x, t) = \max_d (u(x, d, t) + V(p(x, d, t), t + 1)) \quad \text{for } t < T$$

Value function

# Secretary Problem

Also called the **matching or marriage problem**

- ▶ You will see items (spouses) from a distribution of types  $F(x)$ .
- ▶ *You clearly would like to pick the maximum.*  
You see these chronologically.  
After you decide no, you can't go back and select it.
- ▶ **Strategy:** wait for the length of time

$$\frac{1}{e} = \frac{1}{2.718281828} = 0.3678$$

Select after you observe an item greater than the current best.

# Strategy

What's your **best strategy?**

- ▶ Turns out its insensitive to the choice of distribution.
- ▶ Although there is the random sample i.i.d. assumption lurking.
- ▶ You'll not doubt get married between 18 and 60.  
Waiting  $\frac{1}{e}$  along this sequence gets you to the age 32!
- ▶ Then, pick the next best person!

# Q-Learning and Deal-No Deal

Rule of Thumb: Continue as long as there are two large prizes left.

- ▶ Continuation value is large.

For example, with three prizes and two large ones. Risk averse people will naively choose deal, when if they incorporated the continuation value they would choose no deal.

- ▶ Data: Suzanne and Frank's choices on the Dutch version of the show

# Susanne's Choices

Prize	1	2	3	4	5	6	7	8	9
€0.01	x	x	x	x					
€0.20	x	x	x						
€0.50	x	x	x	x	x	x	x		
€1									
€5									
€10									
€20	x	x							
€50	x	x							
€100	x	x	x	x					
€200									
€300	x	x	x						
€400	x								
€500									
€1,000	x	x	x	x	x	x	x		
€2,500	x	x	x	x	x	x	x		
€5,000	x								
€7,500									
€10,000	x	x							
€12,500	x	x	x						
€15,000	x								
€20,000	x	x							
€25,000	x	x	x	x	x				
€50,000	x								
€100,000	x	x	x	x	x	x	x	x	x
€150,000	x	x	x	x	x	x	x	x	x
€250,000	x								
Average €	32,094	21,431	26,491	34,825	46,417	50,700	62,750	83,667	125,000
Offer €	3,400	4,350	10,000	15,600	25,000	31,400	46,000	75,300	125,000
Offer %	11%	20%	38%	45%	54%	62%	73%	90%	100%
Decision	No Deal								

## Table: Frank's Choices

Prize	1	2	3	4	5	6	7	8	9
€0.01	×	×							
€0.20	×	×							
€0.50	×	×	×	×	×	×	×		
€1	×	×	×	×	×				
€5									
€10	×	×	×	×	×	×	×	×	
€20	×	×	×	×	×	×	×	×	
€50									
€100									
€500									
€1,000	×								
€2,500	×	×	×						
€5,000	×	×							
€7,500									
€10,000	×	×	×	×	×	×	×	×	
€25,000	×	×							
€50,000	×	×	×	×					
€75,000	×	×	×						
€100,000	×	×	×						
€200,000	×	×	×		×				
€300,000	×								
€400,000	×								
€500,000	×	×	×	×	×	×	×		
€1,000,000	×								
€2,500,000									
€5,000,000	×								
Average €	383,427	64,502	85,230	95,004	85,005	102,006	2,508	3,343	5,005
Offer €	17,000	8,000	23,000	44,000	52,000	75,000	2,400	3,500	6,000
Offer %	4%	12%	27%	46%	61%	74%	96%	105%	120%
Decision	No Deal								

# Q-Learning

## Rationalising Waiting

There's a matrix of  $Q$ -values that solves the problem.

- ▶ Let  $s$  denote the current state of the system and  $a$  an action.
- ▶ The  $Q$ -value,  $Q_t(s, a)$ , is the value of using action  $a$  today and then proceeding optimally in the future. We use  $a = 1$  to mean no deal and  $a = 0$  means deal.
- ▶ The Bellman equation for  $Q$ -values becomes

$$Q_t(s, a) = u(s, a) + \sum_{s^*} P(s^* | s, a) \max_a Q_{t+1}(s^*, a)$$

# Value Function

The value function and optimal action are given by

$$V(s) = \max_a Q(s, a) \text{ and } a^* = \operatorname{argmax} Q(s, a)$$

- ▶ Transition Matrix.

Consider the problem where you have three prizes left. Now  $s$  is the current state of three prizes.

$$s^* = \{\text{all sets of two prizes}\} \text{ and } P(s^*|s, a = 1) = \frac{1}{3}$$

where the transition matrix is uniform to the next state.

- ▶ There's no continuation for  $P(s^*|s, a = 0)$ .

# Utility

- ▶ Utility.

The utility of the next state depends on the contestants value for money and the bidding function of the banker

$$u(B(s^*)) = \frac{B(s^*)^{1-\gamma} - 1}{1 - \gamma}$$

in power utility case.

# Banker's Function

Expected value implies  $B(s) = \bar{s}$  where  $s$  are the remaining prizes.

- ▶ The website uses the following criteria: with three prizes left:

$$B(s) = 0.305 \times \text{big} + 0.5 \times \text{small}$$

and with two prizes left

$$B(s) = 0.355 \times \text{big} + 0.5 \times \text{small}$$

# Example

Three prizes left:  $s = \{750, 500, 25\}$ .

- ▶ Assume the contestant is risk averse with log-utility  $U(x) = \ln x$ .
- ▶ Banker offers the expected value we get

$$u(B(s = \{750, 500, 25\})) = \ln(1275/3) = 6.052$$

and so  $Q_t(s, a = 0) = 6.052$ .

- ▶ In the continuation problem,  $s^* = \{s_1^*, s_2^*, s_3^*\}$  where  $s_1^* = \{750, 500\}$  and  $s_2^* = \{750, 25\}$  and  $s_3^* = \{500, 25\}$ .

# *Q*-values

We'll have offers 625, 387.5, 137.5 under the expected value.

- ▶ As the banker offers expected value the optimal action at time  $t + 1$  is to take the deal  $a = 0$  with Q-values given by

$$\begin{aligned} Q_t(s, a = 1) &= \sum_{s^*} P(s^* | s, a = 1) \max_a Q_{t+1}(s^*, a) \\ &= \frac{1}{3} (\ln(625) + \ln(387.5) + \ln(262.5)) = 5.989 \end{aligned}$$

as immediate utility  $u(s, a) = 0$ .

- ▶ Hence as

$$Q_t(s, a = 1) = 5.989 < 6.052 = Q_t(s, a = 0)$$

the optimal action is  $a^* = 0$ , deal.

# Utilities

## Continuation value

Not large enough to overcome the generous (expected value) offered by the banker.

- ▶ Sensitivity analysis: Different Banker's bidding function:  
If we use the function from the website (2 prizes):

$$B(s) = 0.355 \times \text{big} + 0.5 \times \text{small}$$

Hence

$$B(s_1^* = \{750, 500\}) = 516.25$$

$$B(s_2^* = \{750, 25\}) = 278.75$$

$$B(s_3^* = \{500, 25\}) = 190$$

# Optimal Action

- ▶ The optimal action with two prizes left for the contestant is

$$Q_{t+1}(s_1^*, a = 1) = \frac{1}{2} (\ln(750) + \ln(500)) = 6.415$$

$$> 6.246 = Q_{t+1}(s_1^*, a = 0) = \ln(516.25)$$

$$Q_{t+1}(s_1^*, a = 1) = \frac{1}{2} (\ln(750) + \ln(25)) = 4.9194$$

$$< 5.63 = Q_{t+1}(s_1^*, a = 0) = \ln(278.75)$$

$$Q_{t+1}(s_1^*, a = 1) = \frac{1}{2} (\ln(500) + \ln(25)) = 4.716$$

$$< 5.247 = Q_{t+1}(s_1^*, a = 0) = (516.25)$$

Hence future optimal policy will be no deal under  $s_1^*$ , and deal under  $s_2^*, s_3^*$ .

# Solve for $Q$ -values

- ▶ Therefore solving for  $Q$ -values at the previous step gives

$$\begin{aligned} Q_t(s, a = 1) &= \sum_{s^*} P(s^* | s, a = 1) \max_a Q_{t+1}(s^*, a) \\ &= \frac{1}{3} (6.415 + 5.63 + 5.247) = 5.764 \end{aligned}$$

with a monetary equivalent as  $\exp(5.764) = 318.62$ .

# Premium

With three prizes we have

$$\begin{aligned}Q_t(s, a = 0) &= u(B(s = \{750, 500, 25\})) \\&= \ln(0.305 \times 750 + 0.5 \times 25) \\&= \ln(241.25) = 5.48.\end{aligned}$$

The contestant is offered \$ 241.

- ▶ Now we have  $Q_t(s, a = 1) = 5.7079 > 5.48 = Q_t(s, a = 0)$  and the optimal action is  $a^* = 1$ , no deal. The continuation value is large.
- ▶ The premium is \$ 241 compared to \$319, a 33% premium.

# Overview

- ▶ Bayes Decision Problems (utility maximization under uncertainty)
- ▶ Reinforcement Learning (Bellman optimality)
- ▶ Imitation Learning (Learn from the best)
- ▶ Q-Learning (Model Free Learning, use DL to approximate the world)

Deep Learning for policy and value functions  
interpolation

# Previous Bayesian Work

- ▶ Müller, Peter, and Giovanni Parmigiani. "Optimal design via curve fitting of Monte Carlo experiments." *Journal of the American Statistical Association* 90.432 (1995): 1322-1330.
- ▶ Müller, Peter. "Simulation based optimal design." *Handbook of Statistics* 25 (2005): 509-518.
- ▶ Moreno-Daz, A., Virto, M. A., Martn, J., & Insua, D. R. Approximate solutions to semi Markov decision processes through Markov chain Montecarlo methods. In *International Conference on Computer Aided Systems Theory* (2003). Springer.
- ▶ Polson, Nicholas G., and Morten Sorensen. "A simulationbased approach to stochastic dynamic programming." *Applied Stochastic Models in Business and Industry* 27.2 (2011): 151-163.
- ▶ Ekin, Tahir, Nicholas G. Polson, and Refik Soyer. "Augmented Markov chain Monte Carlo simulation for two-stage stochastic programs with recourse." *Decision Analysis* 11.4 (2014): 250-264.

# Deep Learning: Kolmogorov-Arnold

There are no multivariate functions just superpositions of univariate ones

Let  $f_1, \dots, f_L$  be given univariate activation functions. We set

$$f_l^{W,b} = f_l \left( \sum_{j=1}^{N_l} W_{lj} X_j + b_l \right) = f_l(W_l X_l + b_l), \quad 1 \leq l \leq L,$$

Our deep predictor has hidden units  $N_l$  and depth  $L$ .

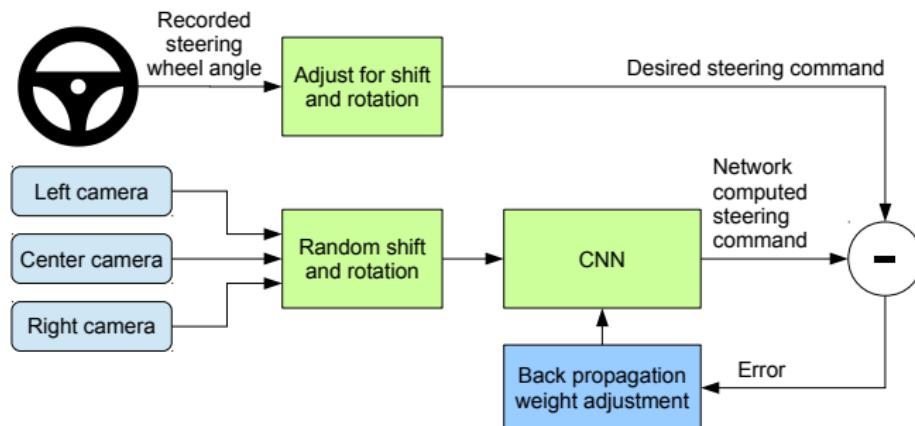
$$\hat{Y}(X) = F(X) = \left( f_1^{W_1, b_1} \circ \dots \circ f_L^{W_L, b_L} \right) (X)$$

Put simply, we model a high dimensional mapping  $F$  via the superposition of univariate semi-affine functions.

Use DL to approximate policy and value functions

# Imitation Learning

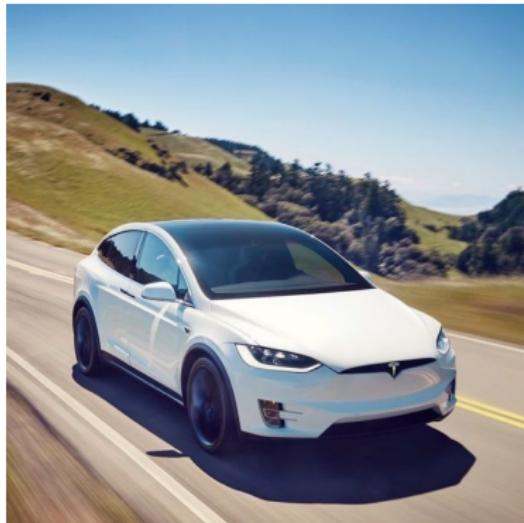
- ▶ Learn from the best. Take all the moves of a chess grand master:  $\{s_i, a_i\}_{i=1}^N$  state-action pairs
- ▶ Learn conditional distribution over actions  $\pi_\theta(a_t | s_t)$
- ▶ Use deep neural network



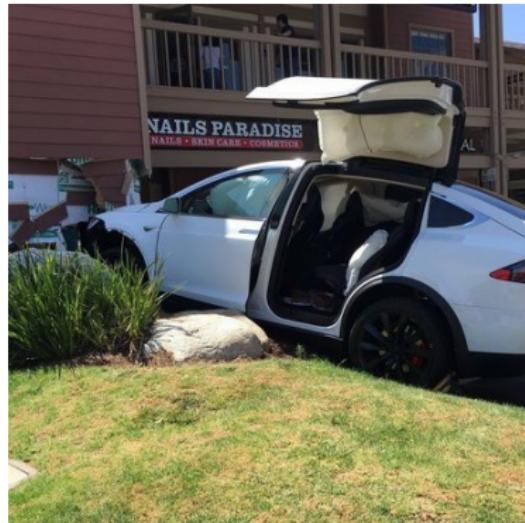
Source: Mariusz 2016

# Reward Function

- ▶ What if some observed actions are bad, e.g in a self-play setting
- ▶ How do we distinguish a good action from a bad action?
- ▶ Introduce the reward function  $r(s, a)$



High Reward



Low Reward

# *Q*-learning

There's a matrix of *Q*-values that solves the problem.

- ▶ Let  $s$  denote the current state of the system and  $a$  an action.
- ▶ The *Q*-value,  $Q_t(s, a)$ , is the value of using action  $a$  today and then proceeding optimally in the future. We use  $a = 1$  to mean no deal and  $a = 0$  means deal.
- ▶ The Bellman equation for *Q*-values becomes

$$Q_t(s, a) = u(s, a) + \sum_{s^*} P(s^*|s, a) \max_a Q_{t+1}(s^*, a)$$

where  $P$  denotes the transition matrix of states

The value function and optimal action are given by

$$V(s) = \max_a Q(s, a) \text{ and } a^* = \operatorname{argmax}_a Q(s, a)$$

# Model-Based RL

- ▶ Stochastic interactions with the environment  $p(s' | s, a)$  is known
- ▶  $p(s' | s, a)$  is an MDP (Markov Decision Process)

$$p_\theta(s^T, a^T) = p(s_1) \prod_{t=1}^T \pi_\theta(a_t | s_t) p(s_{t+1} | s_t, a_t)$$

- ▶ Find policy that maximizes expected reward

$$\theta^* =_{\theta} E_{s^T, a^T} \left[ \sum_{t=1}^T r(s_t, a_t) \right]$$

- ▶  $E_{s^T, a^T}$  is smooth, even when  $s$  or  $a$  are discrete!

# Policy Gradient

- ▶ Specify parametric  $\pi$  and  $r$  (e.g. deep learning)
- ▶ Generate state-action samples  $s_t^i, a_t^i$  and associated reward  $r_t^i$ ,  
 $i = 1, \dots, N$

$$E_{\theta} \left[ \sum_t r_t \right] \approx \frac{1}{N} \sum_{i=1}^N \sum_t r_t^i$$

- ▶ Run a step of batch SGD to update  $\theta$ : policy update via backpropagation, after seeing the reward
- ▶ Only works for deterministic dynamics!
- ▶ Naive algorithm

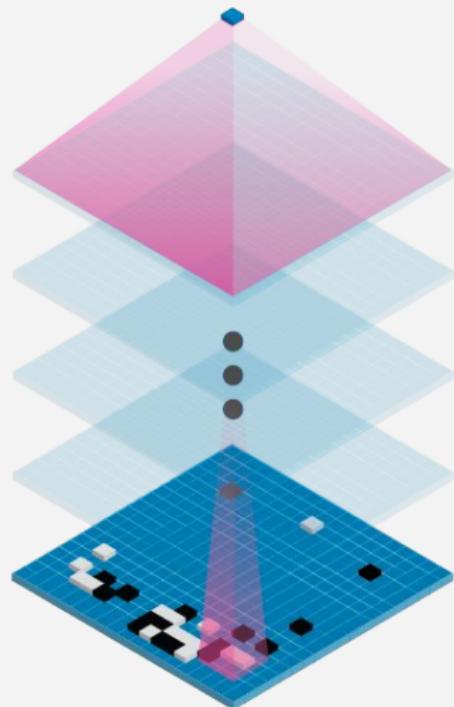
# AlphaGo and AlphaGo Zero

- ▶ Hand-crafted heuristic rules with deep learners
- ▶ Maximize probability of winning (Value function)
- ▶ Use SGD to update network weights based on self-play samples
- ▶ 4 hours to train grand-master level algorithm with no human inputs
- ▶ Same idea can be applied to many other settings: replace models of the world with neural nets
- ▶ Humans do the same. Tennis players do not use Newton's laws to predict trajectory of a ball

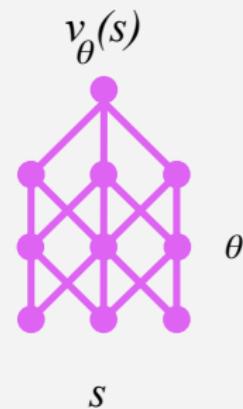
AlphaGo Movie Trailer

# Value Function

Evaluation

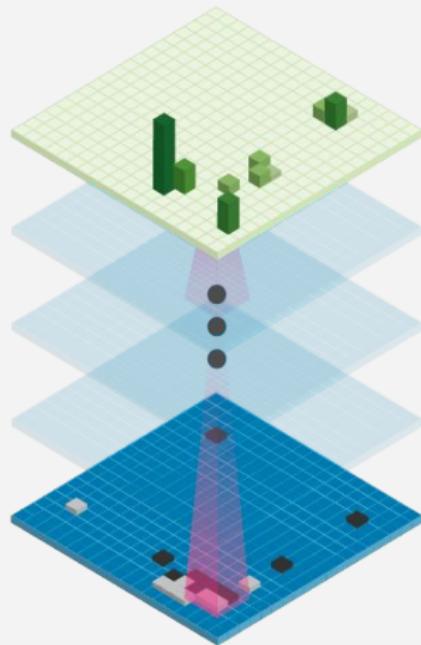


Position



# Policy Function

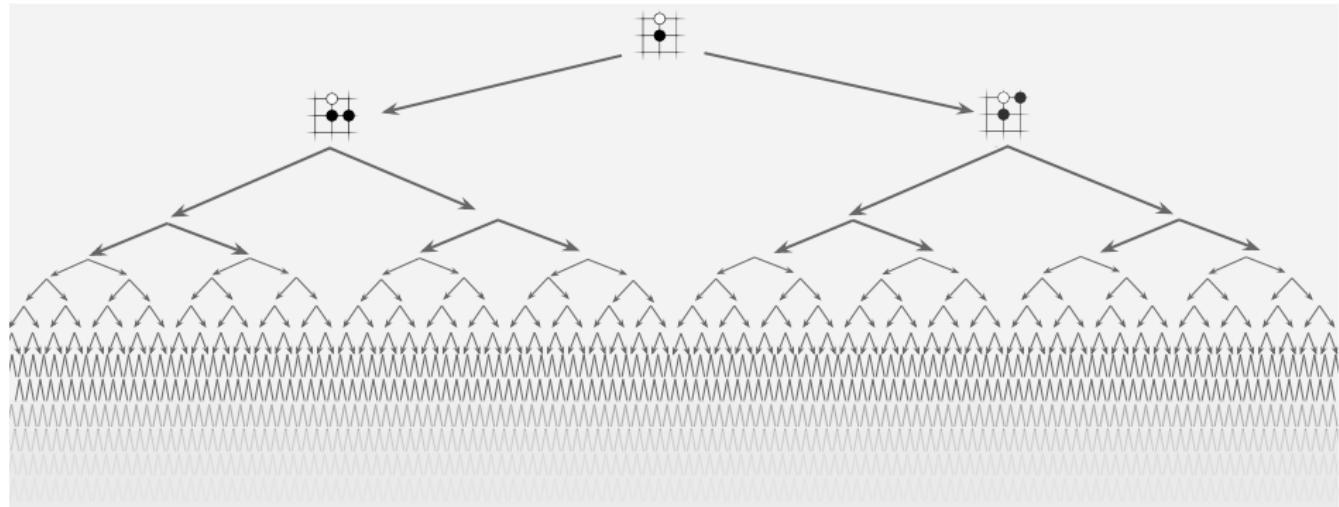
Move probabilities



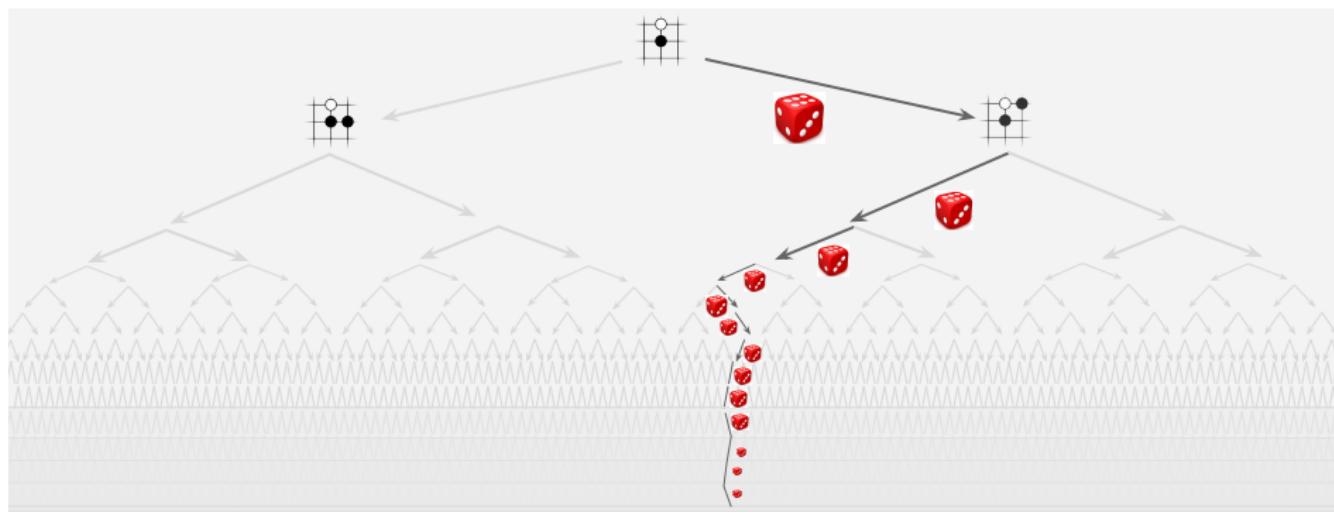
$$p_{\sigma}(a|s)$$

$s$

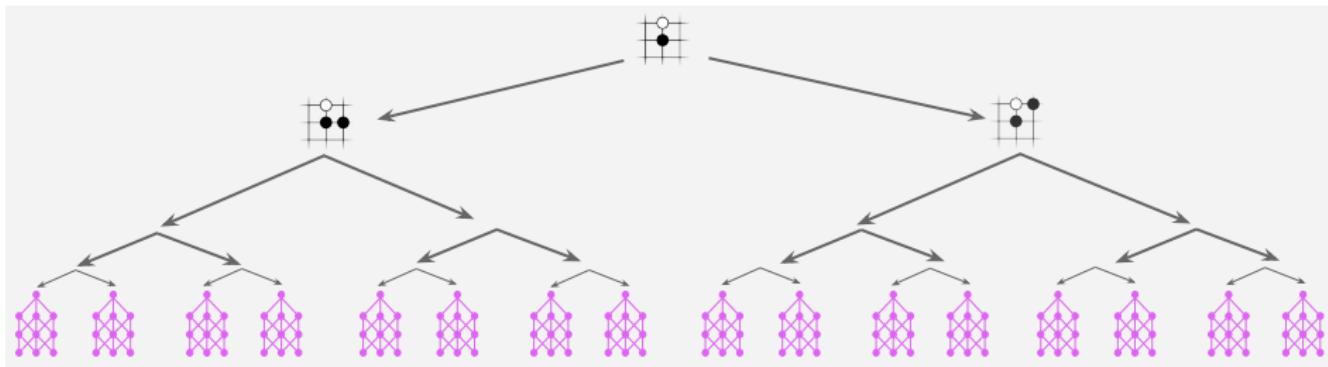
# Full Tree



# Monte-Carlo rollouts

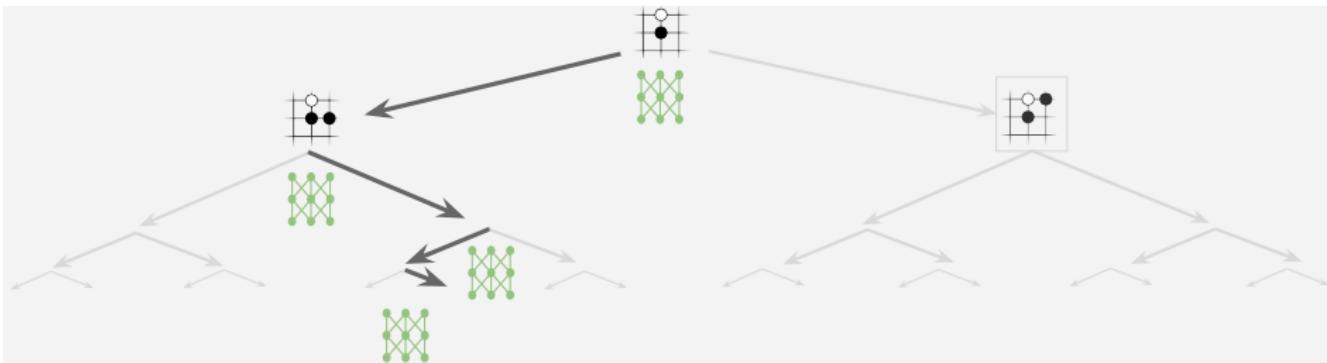


# Reducing depth with value network



- ▶ Value function approximates probability of winning.
- ▶ Pick the path with highest approximated chance to win the game
- ▶ No need to explore the tree till the end

# Reducing breadth with policy network



- ▶ Policy function gives a histogram over possible moves
- ▶ Pick a few with highest probabilities
- ▶ No need to explore low probability moves, reduce breadth of the search

# Deep RL for Portfolio

- ▶ Every  $\Delta t = 30$  minutes re-allocate funds
- ▶ Use  $v_t$  - price vector at the end of period  $t$
- ▶  $v_t^h$  and  $v_t^l$  are high and low proces for the period  $t$
- ▶ Relative return is  $y_t = v_t / v_{t-1}$
- ▶  $p_t = p_{t-1} y_t w_{t-1}$  is portfolio value at the end of period  $t$

The goal of portfolio manager it to re-calculate weights  $w_t \rightarrow w_{t+1}$ , given  $X_t = (y^t, y^{th}, y^{tl})$ ,  $y^t = (y_{t-n+1}, \dots, y_t)$

# Deep RL for Portfolio

The goal is to maximize the final portfolio return  $R$ , as a function of states  $s_t = (X_t, w_{t-1})$  and actions  $a_t = w_t$

$$R(s_1, a_1, \dots, s_{t_f}, a_{t_f}, s_{t_f+1}) = \frac{1}{t_f} \ln \frac{p_f}{p_0} = \frac{1}{t_f} \sum_{t=1}^{t_f+1} \ln (\mu_t y_t \star w_{t-1}) \quad (1)$$

$$= \frac{1}{t_f} \sum_{t=1}^{t_f+1} r_t. \quad (2)$$

$$p_f = p_0 \exp \left( \sum_{t=1}^{t_f+1} r_t \right) = p_0 \prod_{t=1}^{t_f+1} y_t \star w_{t-1}$$

"

# Policy Gradient

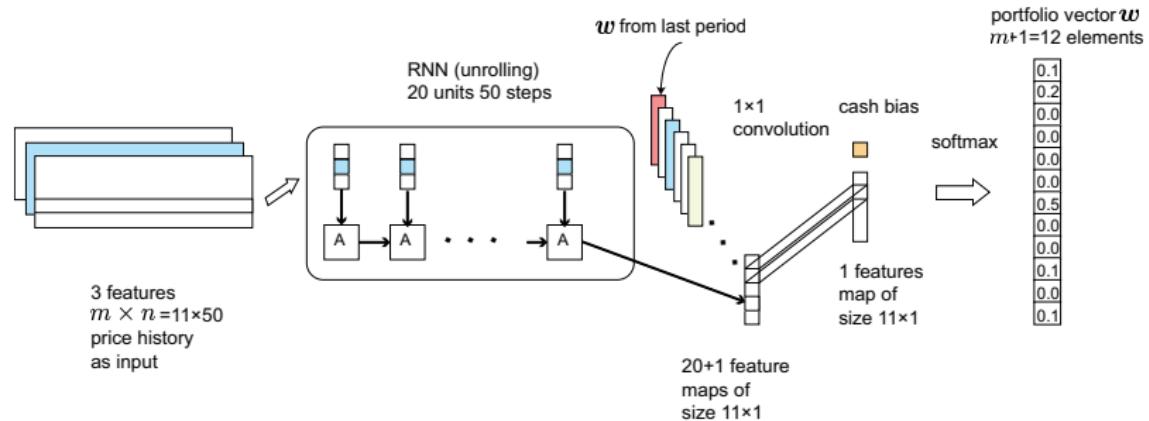
- ▶ Goal is to find an optimal policy  $\pi_\theta : \mathcal{S} \rightarrow \mathcal{A}$ , here  $\pi_\theta$  is a deep learner
- ▶ Optimize the reward

$$J(\pi_\theta) = R(s_1, \pi_\theta(s_1), \dots, s_{t_f}, \pi_\theta(s_{t_f}), s_{t_f+1})$$

- ▶ Use SGD

$$\theta \longrightarrow \theta + \lambda \nabla_\theta J_{[0,t_f]}(\pi_\theta).$$

# Network Architecture



# Empirical Results

Five stock portfolio: APPL, V, BABA, ADBE, SNE



# Discussion

- ▶ Merging RL with previous Bayesian decision work with lots of applications
- ▶ Computational bottlenecks have been solved by DL
- ▶ Nature paper: Mastering the game of Go without human knowledge
- ▶ Reinforcement Learning notes by David Silver
- ▶ Open area of research

# Overview

Bayes accounts for probability in many scenarios

- ▶ Prisoner's Dilemma
- ▶ Monte Hall.
- ▶ Island Problem
- ▶ Sally Clark
- ▶ O.J. Simpson
- ▶ Bellman and Sequential Learning

# Probability: 41901

## **Week 3: Distributions and Expectations**

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Topics

1. Markov Dependence
2. SP500 Hidden Markov Model
3. Random Variables
4. Transformations
5. Expectations
6. Moment Generating Functions

Reading: DeGroot and Schervish: 3, p.97-181 and 4, 181-247

# Markov Dependence

- We can always factor a joint distribution as

$$p(X_n, X_{n-1}, \dots, X_1) = p(X_n | X_{n-1}, \dots, X_1) \dots p(X_2 | X_1) p(X_1)$$

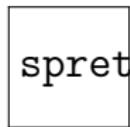
## Example

- A process has the *Markov Property* if

$$p(X_n | X_{n-1}, \dots, X_1) = p(X_n | X_{n-1})$$

- Only the current history matter when determining the probabilities.

# SP500 daily ups and downs



spreturns-new.jpg

- ▶ Daily return data from 1948 to 2007 for the SP500 index of stocks
- ▶ Can we calculate the probability of ups and downs?

# A real world probability model

## Hidden Markov Models

Are stock returns a random walk?

Hidden Markov Models (Baum-Welch, Viterbi)

- ▶ Daily returns on the SP500 stock market index.  
Build a hidden Markov model to predict the ups and downs.
- ▶ Suppose that stock market returns on the next four days are  $X_1, \dots, X_4$ .
- ▶ Let's empirical determine conditionals and marginals

# SP500 Data

## Marginal and Bivariate Distributions

- ▶ Empirically, what do we get? Daily returns from 1948 – 2007.

$x$	Down	Up
$P(X_i) = x$	0.474	0.526

- ▶ Finding  $p(X_2|X_1)$  is twice as much computational effort: counting  $UU, UD, DU, DD$  transitions.

$X_i$	Down	Up
$X_{i-1} = Down$	0.519	0.481
$X_{i-1} = Up$	0.433	0.567

# Conditioned on two days

- ▶ Let's do  $p(X_3|X_2, X_1)$

$X_{i-2}$	$X_{i-1}$	Down	Up
Down	Down	0.501	0.499
Down	Up	0.412	0.588
Up	Down	0.539	0.461
Up	Up	0.449	0.551

- ▶ We could do the distribution  $p(X_2, X_3|X_1)$ . This is a joint, marginal and conditional distribution all at the same time.  
*Joint* because more than one variable ( $X_2, X_3$ ), *marginal* because it ignores  $X_4$  and *conditional* because its given  $X_1$ .

# Joint Probabilities

- ▶ Under Markov dependence

$$\begin{aligned}P(UUD) &= p(X_1 = U)p(X_2 = U|X_1 = U)p(X_3|X_2 = U, X_1 = U) \\&= (0.526)(0.567)(0.433)\end{aligned}$$

- ▶ Under independence we would have got

$$\begin{aligned}P(UUD) &= P(X_1 = U)p(X_2 = U)p(X_3 = D) \\&= (.526)(.526)(.474) \\&= 0.131\end{aligned}$$

# Continuous Random Variables

1. We will be interesting in random variables  $X : \Omega \rightarrow \mathfrak{N}$
2. The *cumulative distribution function* of  $X$  is defined as

$$F_X(x) = \mathbb{P}(X \leq x)$$
$$1 - F_X(x) = \mathbb{P}(X > x)$$

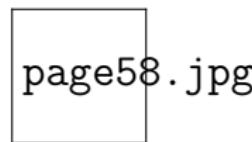
3. The *probability density function* of  $X$  is defined as

$$f_X(x) = F'_X(x) \text{ and } \mathbb{P}(X \leq x) = \int_{-\infty}^x f_X(x) dx$$

4. We have the following properties:  $f_X(x) \geq 0$  and  $F_X(x)$  is non-decreasing
5.  $\lim_{x \rightarrow -\infty} F_X(x) = 0$  and  $\lim_{x \rightarrow \infty} F_X(x) = 1$ .

# Intuition: cdf

- ▶ For a discrete random variable  $F_X(x)$  is a step function with the heights of the steps being  $\mathbb{P}(X = x_i)$



- ▶ For a continuous variable it's a continuous non-decreasing curve.

# Expectation and Variance

- ▶ The *expectation* of  $X$  is given by

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} xf_X(x)dx$$

- ▶ The *variance* of  $X$  is given by

$$Var(X) = \mathbb{E}((X - \mu)^2) = \mathbb{E}(X^2) - \mu^2$$

- ▶ Also derivatives of moment generating functions (mgfs).

# Intuition: pdf

- ▶ The intuitive interpretation of a pdf is that for small  $\Delta x$

$$\mathbb{P}(x < X \leq x + \Delta x) = \int_x^{x+\Delta x} f_X(x) dx \approx f_X(x)\Delta x$$

X lies in a small interval with probability proportional to  $f_X(x)$ .



- ▶ We will also consider the multivariate generalization of this.

# Transformations

The cdf identity gives

$$F_Y(y) = \mathbb{P}(Y \leq y) = \mathbb{P}(g(X) \leq y)$$

- ▶ Hence if the function  $g(\cdot)$  is monotone we can invert to get

$$F_Y(y) = \int_{g(x) \leq y} f_X(x) dx$$

- ▶ If  $g$  is increasing  $F_Y(y) = P(X \leq g^{-1}(y)) = F_X(g^{-1}(y))$   
If  $g$  is decreasing  $F_Y(y) = P(X \geq g^{-1}(y)) = 1 - F_X(g^{-1}(y))$

# Transformation Identity

1. Theorem 1: Let  $X$  have pdf  $f_X(x)$  and let  $Y = g(X)$ . Then if  $g$  is a monotone function we have

$$f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$$

There's also a multivariate version of this that we'll see later.

- ▶ Suppose  $X$  is a continuous rv, what's the pdf for  $Y = X^2$ ?
- ▶ Let  $X \sim N(0, 1)$ , what's the pdf for  $Y = X^2$ ?

# Probability Integral Transform

## Theorem

Suppose that  $U \sim U[0, 1]$ , then for any continuous distribution function  $F$ , the random variable  $X = F^{-1}(U)$  has distribution function  $F$ .

- ▶ Remember that for  $u \in [0, 1]$ ,  $\mathbb{P}(U \leq u) = u$ , so we have

$$\mathbb{P}(X \leq x) = \mathbb{P}\left(F^{-1}(U) \leq x\right) = \mathbb{P}(U \leq F(x)) = F(x)$$

Hence,  $X = F_X^{-1}(U)$ .

# Inequalities and Identities

## 1. Markov

$$\mathbb{P}(g(X) \geq c) \leq \frac{\mathbb{E}(g(X))}{c} \text{ where } g(X) \geq 0$$

## 2. Chebyshev

$$\mathbb{P}(|X - \mu| \geq c) \leq \frac{\text{Var}(X)}{c^2}$$

## 3. Jensen

$$\mathbb{E}(\phi(X)) \leq \phi(\mathbb{E}(X))$$

## 4. Cauchy-Schwarz

$$\text{corr}(X, Y) \leq 1$$

Chebyshev follows from Markov. Mike Steele and Cauchy-Schwarz.

# Conditional Expectation

Let  $X$  denote our random variable and  $Y$  a conditioning variable.

- ▶ In the continuous case

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

- ▶ The conditional expectation (given  $Y$ ) is determined by

$$\mathbb{E}(g(X)|Y) = \int_{-\infty}^{\infty} g(x)f_X(x|Y)dx$$

where  $f_{X|Y}(x|y)$  denotes the conditional density of  $X|Y$ .

# Iterated Expectation

- ▶ The following fact is known as the law of iterated expectation

$$\mathbb{E}_Y \left( \mathbb{E}_{X|Y} (X|Y) \right) = \mathbb{E} (X)$$

This is a very useful way of calculating  $E (X)$  (martingales)

- ▶ Sometimes called the *Tower property of expectations*.
- ▶ A similar decomposition for a variance

$$Var(X) = \mathbb{E} (Var(X|Y)) + Var (\mathbb{E}(X|Y)) .$$

# Exponential Distribution

## Example

- One of the most important distributions is the *exponential*

$$f_X(x) = \lambda e^{-\lambda x} \text{ and } F_X(x) = \int_0^x \lambda e^{-\lambda x} dx = 1 - e^{-\lambda x}$$

$\lambda$  is called a parameter.

- The *expectation* of X is

$$\mathbb{E}(X) = \int_0^\infty x \lambda e^{-\lambda x} dx = \int_0^\infty x d(-e^{-\lambda x}) = \frac{1}{\lambda}$$

- The *variance* of X is

$$Var(X) = \mathbb{E}(X^2) - \mu^2 = \frac{1}{\lambda^2}$$

# Simulation

Suppose that  $X \sim F_X(x)$  and let  $Y = g(X)$ .  
How do we find  $F_Y(y)$  and  $f_Y(y)$  ?

- ▶ von Neumann

Given a uniform  $U$ , how do we find  $X = g(U)$ ?

- ▶ In the bivariate case  $(X, Y) \rightarrow (U, V)$ .

We need to find  $f_{(U,V)}(u, v)$  from  $f_{X,Y}(x, y)$

- ▶ Applications: Simulation, MCMC and PF.

# Transformations

An important application is how to transform multiple random variables?

Here's the basic set-up:

- ▶ Suppose that we have random variables:

$$(X, Y) \sim f_{X,Y}(x, y)$$

A transformation of interest given by:

$$U = g(X, Y) \text{ and } V = h(X, Y)$$

- ▶ The problem is how to compute  $f_{U,V}(u, v)$ ? Jacobian

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix}$$

# Bivariate Change of Variable

- ▶ *Theorem:* (change of variable)

$$f_{U,V}(u,v) = f_{X,Y}(h_1(u,v), h_2(u,v)) \left| \frac{\partial(x,y)}{\partial(u,v)} \right|$$

The last term is the Jacobian.

This can be calculated in two ways.

$$\left| \frac{\partial(x,y)}{\partial(u,v)} \right| = 1 / \left| \frac{\partial(u,v)}{\partial(x,y)} \right|$$

- ▶ So we don't always need the inverse transformation  
 $(x,y) = (g^{-1}(u,v), h^{-1}(u,v))$

# Example: Exponentials

## Example

- ▶ Suppose that  $X$  and  $Y$  are independent, identically distributed random variables each with an  $\text{Exp}(\lambda)$  distribution. Let

$$U = X + Y \text{ and } V = \frac{X}{X + Y}$$

- ▶ The joint probability distribution

$$f_{X,Y}(x,y) = \lambda^2 e^{-\lambda(x+y)} \text{ for } 0 < x, y < \infty$$

- ▶ The inverse transformation is  $x = uv, y = u(1 - v)$ .  
The range of definition  $0 < u < \infty, 0 < v < 1$ .

# Jacobian

- We can calculate the Jacobian as

$$J = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} = \begin{vmatrix} v & u \\ 1-v & -u \end{vmatrix}$$

- Thus the joint density of  $(U, V)$  is

$$f_{U,V}(u, v) = \lambda^2 u e^{-\lambda u}$$

- Therefore

$$U \sim \Gamma(2, \lambda) \text{ and } V \sim U(0, 1)$$

# Example: Normals

Suppose that  $X, Y$  are i.i.d. each with a standard normal  $N(0, 1)$  distribution.

Let  $D = X^2 + Y^2$  and  $\Theta = \tan^{-1} \left( \frac{Y}{X} \right)$ .

Joint distribution

$$f_{X,Y}(x,y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \frac{1}{\sqrt{2\pi}} e^{-\frac{y^2}{2}} \text{ for } -\infty < x, y < \infty$$

We can calculate the Jacobian as

$$J = \begin{vmatrix} \frac{\partial d}{\partial x} & \frac{\partial d}{\partial y} \\ \frac{\partial \theta}{\partial x} & \frac{\partial \theta}{\partial y} \end{vmatrix} = \begin{vmatrix} 2x & 2y \\ -\frac{y}{x^2+y^2} & \frac{x}{x^2+y^2} \end{vmatrix}$$

# Example: Normal Simulation

- ▶ Thus the joint density of  $(D, \Theta)$  is

$$f_{D,\Theta}(d, \theta) = \frac{1}{4\pi} e^{-\frac{d}{2}} \text{ where } 0 \leq d < \infty, 0 \leq \theta \leq 2\pi$$

- ▶ Therefore, we have

$$D \sim \text{Exp}\left(\frac{1}{2}\right) \text{ and } \Theta \sim U(0, 2\pi)$$

- ▶ Box-Muller transform for simulating normal random variables:  
draw  $D = -2 \log U_1$  and  $\Theta = 2\pi U_2$

$$X = \sqrt{D} \cos(\Theta) = \sqrt{-2 \log U_1} \cos(2\pi U_2)$$

$$Y = \sqrt{D} \sin(\Theta) = \sqrt{-2 \log U_1} \sin(2\pi U_2)$$

# Cauchy: ratio of normals: $C = X/Y$

Let  $X, Y \sim N(0, 1)$ .

$$f_{X,Y}(x,y) = \frac{1}{2\pi} e^{-\frac{1}{2}(x^2+y^2)} \quad -\infty < x, y < \infty$$

Transformation  $U = X/Y$  and  $V = |Y|$ . Inverse  $x = uv, y = v$   
Jacobian

$$\begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

Joint and marginal

$$f_{U,V}(u,v) = \frac{1}{2\pi} v e^{-\frac{1}{2}v^2(1+u^2)} \quad 0 < v < \infty, -\infty < u < \infty.$$

$$f_U(u) = \int_v f_{U,V}(u,v) dv = 2 \int_0^\infty \frac{1}{2\pi} v e^{-\frac{1}{2}v^2(1+u^2)} dv = \frac{1}{\pi} \frac{1}{1+u^2}$$

A Cauchy distribution.

# Special Distributions

See *Common Distributions*

1. Bernoulli and Binomial
  2. Hypergeometric
  3. Poisson
  4. Negative Binomial
  5. Normal Distribution
  6. Gamma Distribution
  7. Beta Distribution
  8. Multinomial Distribution
  9. Bivariate Normal Distribution
  10. Wishart Distribution
- ...

# Binomial Distribution

This models the number of heads occurring in  $n$  successive trials of the Bernoulli coin.

- ▶ Thus  $\Omega = \{0, 1, \dots, n\}$  and

$$f_X(x) = \binom{n}{x} p^x (1-p)^{n-x} \quad 0 \leq x \leq n$$

- ▶ Generalizations?  
Random probability and hierarchical exchangeable models.

# Sums of Binomial Distributions

## Distribution of sums

- ▶ The probability generating function is

$$\mathbb{E}(z^X) = \sum_{r=0}^n z^r \binom{n}{r} p^r q^{n-r} = (pz + q)^n$$

- ▶ Therefore  $X + Y \sim \text{Bin}(n + m, p)$
- ▶ In general  $X_i \sim \text{Bin}(n_i, p)$  then  $\sum_{i=1}^k X_i \sim \text{Bin}\left(\sum_{i=1}^k n_i, p\right)$

# Sums of Poisson Distributions

$X + Y$  where  $X \sim Poi(\lambda)$  and  $Y \sim Poi(\mu)$ ?

- ▶ The probability generating function

$$\mathbb{E}(z^X) = \sum_{r=0}^{\infty} z^r e^{-\lambda} \frac{\lambda^r}{r!} = e^{-\lambda(1-z)}$$

Let  $z = e^t$  to get the moment generating function.

- ▶ Hence for the sum  $X + Y$  we have

$$e^{-\lambda(1-z)} e^{-\mu(1-z)} = e^{-(\lambda+\mu)(1-z)}$$

Hence  $X + Y \sim Poi(\lambda + \mu)$ .

# Sums of Random Variables

- ▶ For any random variables, the expectation of the sum is the sum of expectations

$$\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}(X_1) + \dots + \mathbb{E}(X_n)$$

- ▶ If  $X_1, \dots, X_n$  are independent and identically distributed (i.i.d)

$$Var\left(\frac{X_1 + \dots + X_n}{n}\right) = \frac{1}{n}Var(X_1)$$

# Cauchy Distribution

The Cauchy distribution has a pdf given by

$$f_X(x) = \frac{1}{\pi} \frac{1}{1+x^2}$$

Special case of a  $t_\nu$ -distribution with density

$$f_X(x|\nu) = C\nu \left(1 + \frac{x^2}{\nu}\right)^{-\frac{1}{2}(1+\nu)}$$

- ▶  $\nu = 1$  then  $T$  is a Cauchy
- ▶  $\nu \rightarrow \infty$  then  $T \rightarrow_d N(0, 1)$
- ▶  $E(T) = 0, Var(T) = \nu / (\nu - 2)$  for  $\nu > 2$ .

# Sums of Cauchy's

## Example

- If  $X_1, \dots, X_n \sim C(0, 1)$  then

$$\frac{1}{n} (X_1 + \dots + X_n) \sim C(0, 1)$$

- Averaging a set of Cauchy's leads back to the same Cauchy !

The normal distribution behaves differently: if  $X_i \sim N(\mu, \sigma^2)$

$$\frac{1}{n} (X_1 + \dots + X_n) \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Eventually  $\bar{X} \rightarrow \mu$  a constant (CLT)

# Moment Generating Functions

- ▶ Definition. The moment generating function (mgf) is defined by  $m_X(t) = E(e^{tX})$  that is

$$m_X(t) = \int_0^{\infty} e^{tX} f_X(x) dx$$

- ▶ Here  $t$  is a parameter.

You'll need to check when the function  $m_X(t)$  exists.

# Moment Identity

- ▶ There is an equivalence between knowing the mgf and the set of moments of the distribution given by the following

*Theorem:* If  $X$  has mgf  $m_X(t)$  then  $\mathbb{E}(X^n) = m_X^{(n)}(0)$

Here  $m_X^{(n)}(0)$  denotes the  $n$ th derivative evaluated at zero.

- ▶ What about a  $C(0, 1)$  random variable ?

# Characteristic Functions

Characteristic Function:  $\phi_X(t)$ .

- ▶ This function exists for all distributions and all values of  $t$ .

It is defined by

$$\phi_X(t) = E\left(e^{itX}\right)$$

- ▶ Here  $e^{itX} = \cos(tX) + i \sin(tX)$  with  $i^2 = -1$ .
- ▶ Technically better to do everything in terms of characteristic functions as they always exist.

$$|\phi_X(t)| \leq \mathbb{E}\left(|e^{itX}| \right) = 1$$

# $\phi_C(t)$ ?

Cauchy distribution  $\phi_C(t) = e^{-|t|}$ .

- ▶ mgf only exists at the origin
- ▶ We can calculate

$$\begin{aligned}\phi_X(t) &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{e^{itx}}{1+x^2} dx \\ &= e^{-|t|}\end{aligned}$$

- ▶ Cauchy's residue theorem calculates the complex integral

# Properties of mgfs

## Theorem

The mgf  $m_X(t) = \mathbb{E}(e^{tX})$  uniquely determines the distribution of  $X$  provided it is defined for some open interval of  $t$  values.

$$m^{(r)}(0) = \mathbb{E}(X^r)$$

- Key identity is

$$\frac{d}{dt}m_X(t) = \mathbb{E}\left(\frac{d}{dt}e^{tX}\right) = \mathbb{E}(Xe^{tX})$$

# Distribution of $\bar{X}$

Let  $(X_1, \dots, X_n)$  be an iid sample from any distribution, then

$$m_{\bar{X}}(t) = m_X \left( \frac{t}{n} \right)^n$$

- ▶ *Example:* Suppose the  $X'_i$ 's are  $N(\mu, \sigma^2)$ . Then

$$m_{\bar{X}}(t) = \exp \left( \mu t + \frac{1}{2} \frac{\sigma^2}{n} t^2 \right)$$

The distribution of  $\bar{X}$  is normal

- ▶ *Example:* Suppose the  $X'_i$ 's are  $\Gamma(\alpha, \beta)$  then  $\bar{X} \sim \Gamma(n\alpha, \beta/n)$ .

# Central Limit Theorem

Let  $X_1, X_2, \dots$  be a random sample (iid) with  $E(X_i) = \mu$  and  $\text{Var}(X_i) = \sigma^2 < \infty$ . Suppose that the mgf exists in a neighborhood of zero ie  $m_X(\theta)$  exists for some  $\theta > 0$ .

## Theorem

*The sample mean converges in distribution to a normal random variable*

$$\lim_{n \rightarrow \infty} \mathbb{P} \left( \left| \frac{X_1 + \dots + X_n - n\mu}{\sigma\sqrt{n}} \leq x \right| \right) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{y^2}{2}} dy = \Phi(x)$$

which is the distribution function of a standard  $N(0, 1)$ .

- ▶ One can expand the mgf  $m_{\bar{X}}(t)$  and take limits as  $n \rightarrow \infty$ .
- ▶  $\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \rightarrow_D Z$

# Sketch Proof

*Key facts: Taylor Series expansion of mgf*



$$M_{\sqrt{n}(\bar{X}_n - \mu)/\sigma}(t) = \left( M_Y \left( \frac{t}{\sqrt{n}} \right) \right)^n$$



$$\left( M_Y \left( \frac{t}{\sqrt{n}} \right) \right)^n = 1 + \frac{1}{2} \left( \frac{t}{\sqrt{n}} \right)^2 + R_Y \left( \frac{t}{\sqrt{n}} \right)$$



$$\lim_{n \rightarrow \infty} \left( M_Y \left( \frac{t}{\sqrt{n}} \right) \right)^n = \exp(t^2/2)$$

This is the mgf of a  $Z \sim N(0, 1)$

# Order Statistics

## *Order Statistics*

- ▶ The order statistics of a random sample  $X_1, \dots, X_n$  are the sample values in ascending order, denoted by  $X_{(1)}, \dots, X_{(n)}$  or equivalently

$$X_{(1)} = \min_{1 \leq i \leq n} X_i = X_{\min} \text{ and } X_{(n)} = \max_{1 \leq i \leq n} X_i = X_{\max}$$

- ▶ The median  $X_{med}$  is also widely studied as it is a consistent estimator for the location of a distribution for a wide family and is less sensitive to extreme observations (breakdown point).

# Proof

Consider  $X_{(1)}$  and  $X_{(n)}$ .

$$\begin{aligned}F_{X_{(n)}}(x) &= P\left(\max_{1 \leq i \leq n} X_i \leq x\right) \\&= (F_X(x))^n\end{aligned}$$

Differentiating with respect to  $x$  gives

$$f_{X_{(n)}}(x) = nf_X(x) (F_X(x))^{n-1}$$

For the minimum, let's compute

$$\begin{aligned}1 - F_{X_{(1)}}(x) &= P(X_{\min} \geq x) \\&= (1 - F_X(x))^n\end{aligned}$$

Differentiating with respect to  $x$  gives

$$f_{X_{(n)}}(x) = nf_X(x) (1 - F_X(x))^{n-1}$$

# General Case

*Theorem:* Let  $X_{(1)}, \dots, X_{(n)}$  denote the order statistics from a population with cdf  $F_X(x)$  and pdf  $f_X(x)$ . Then the pdf of  $X_{(j)}$  is

$$f_{X_{(j)}}(x) = \frac{n}{(j-1)(n-j)} f_X(x) (F_X(x))^{j-1} (1 - F_X(x))^{n-j}$$

- ▶ For the general case we get

$$F_{X_{(j)}}(x) = \sum_{k=j}^n \binom{n}{k} (F_X(x))^k (1 - F_X(x))^{n-k}$$

# Probability: 41901

## **Week 4: Bayesian Statistics**

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Introduction to Bayesian Methods

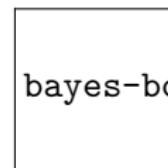
## Modern Statistical Learning

- ▶ Bayes Rule and Probabilistic Learning
- ▶ Computationally challenging: MCMC and Particle Filtering
- ▶ Many applications in Finance:  
Asset pricing and corporate finance problems.

Lindley, D.V. *Making Decisions*  
Bernardo, J. and A.F.M. Smith *Bayesian Theory*

# Bayesian Books

- ▶ Hierarchical Models and MCMC
- ▶ Bayesian Nonparametrics  
Machine Learning
- ▶ Dynamic State Space  
Models ...



# Popular Books

McGrayne (2012): The Theory  
that would not Die

- ▶ History of Bayes-Laplace
- ▶ Code breaking
- ▶ Bayes search: Air France

...

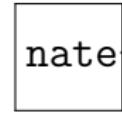


bayes-book.jpg

# Nate Silver: 538 and NYT

Silver (2012): The Signal  
and The Noise

- ▶ Presidential Elections
- ▶ Bayes dominant methodology
- ▶ Predicting College Basketball/Oscars ...



nate-silver-book.jpg

# Things to Know

Explosion of Models and [Algorithms](#) starting in 1950s

- ▶ Bayesian Regularisation and Sparsity
- ▶ Hierarchical Models and Shrinkage
- ▶ Hidden Markov Models
- ▶ Nonlinear Non-Gaussian State Space Models

[Algorithms](#)

- ▶ Monte Carlo Method (von Neumann and Ulam, 1940s)
- ▶ Metropolis-Hastings (Metropolis, 1950s)
- ▶ Gibbs Sampling (Geman and Geman, Gelfand and Smith, 1980s)
- ▶ Sequential Particle Filtering

# Probabilistic Reasoning

## Bayesians only make Probability statements

- ▶ Bayesian Probability (Ramsey, 1926, de Finetti, 1931)
  1. Beta-Binomial Learning: Black Swans
  2. Elections: Nate Silver
  3. Baseball: Kenny Lofton and Derek Jeter
- ▶ Monte Carlo (von Neumann and Ulam, Metropolis, 1940s)  
Shrinkage Estimation: (Lindley and Smith, Efron and Morris, 1970s)

# Bayes Learning: Beta-Binomial

*How do I update my beliefs about a coin toss?*

Likelihood for Bernoulli

$$p(y|\theta) = \prod_{t=1}^T p(y_t|\theta) = \theta^{\sum_{t=1}^T y_t} (1-\theta)^{T - \sum_{t=1}^T y_t}.$$

Initial prior distribution  $\theta \sim \mathcal{B}(a, A)$  given by

$$p(\theta|a, A) = \frac{\theta^{a-1} (1-\theta)^{A-1}}{B(a, A)}$$

# Bayes Learning: Beta-Binomial

Updated posterior distribution is also Beta

$$p(\theta|y) \sim \mathcal{B}(a_T, A_T) \text{ and } a_T = a + \sum_{t=1}^T y_t, A_T = A + T - \sum_{t=1}^T y_t$$

The posterior mean and variance are

$$E[\theta|y] = \frac{a_T}{a_T + A_T} \text{ and } var[\theta|y] = \frac{a_T A_T}{(a_T + A_T)^2 (a_T + A_T + 1)}$$

# Black Swans

Taleb, The Black Swan: the Impact of the Highly Improbable

Suppose you're only see a sequence of White Swans, having never seen a Black Swan.

What's the Probability of Black Swan event *sometime* in the future?

Suppose that after  $T$  trials you have only seen successes  $(y_1, \dots, y_T) = (1, \dots, 1)$ . The next trial being a success has

$$p(y_{T+1} = 1 | y_1, \dots, y_T) = \frac{T+1}{T+2}$$

For large  $T$  is almost certain. Here  $a = A = 1$ .

# Black Swans

*Principle of Induction (Hume)*

The probability of **never** seeing a Black Swan is given by

$$p(y_{T+1} = 1, \dots, y_{T+n} = 1 | y_1, \dots, y_T) = \frac{T+1}{T+n+1} \rightarrow 0$$

**Black Swan** will eventually happen – don't be surprised when it actually happens.

# Bayesian Inference

**Key Idea:** Explicit use of probability for summarizing uncertainty.

1. A **probability distribution** for data given parameters

$$f(y|\theta) \text{ Likelihood}$$

2. A **probability distribution** for unknown parameters

$$p(\theta) \text{ Prior}$$

3. Inference for unknowns conditional on observed data

Inverse probability (Bayes Theorem);

Formal decision making (Loss, Utility)

# Posterior Inference

Bayes theorem to derive posterior distributions

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$
$$p(y) = \int p(y|\theta)p(\theta)d\theta$$

Allows you to make probability statements

- ▶ They can be very different from p-values!  
Hypothesis testing and Sequential problems
- ▶ Markov chain Monte Carlo (MCMC) and Filtering (PF)

# Conjugate Priors

## Example

- ▶ **Definition:** Let  $\mathcal{F}$  denote the class of distributions  $f(y|\theta)$ .  
A class  $\Pi$  of prior distributions is **conjugate** for  $\mathcal{F}$  if the posterior distribution is in the class  $\Pi$  for all  
 $f \in \mathcal{F}, \pi \in \Pi, y \in \mathcal{Y}$ .
- ▶ **Example: Binomial/Beta:**  
Suppose that  $Y_1, \dots, Y_n \sim Ber(p)$ .  
Let  $p \sim Beta(\alpha, \beta)$  where  $(\alpha, \beta)$  are known hyper-parameters.  
The beta-family is very flexible  
Prior mean  $E(p) = \frac{\alpha}{\alpha+\beta}$ .

# Lindley's Paradox

*Often evidence which, for a Bayesian statistician, strikingly supports the null leads to rejection by standard classical procedures.*

- ▶ Do Bayes and Classical always agree?  
Bayes computes the probability of the null being true given the data  $p(H_0|D)$ . That's not the p-value. Why?
- ▶ Surely they agree asymptotically?
- ▶ How do we model the prior and compute likelihood ratios  $L(H_0|D)$  in the Bayesian framework?

# Bayes $t$ -ratio

Edwards, Lindman and Savage (1963)

Simple approximation for the likelihood ratio.

$$L(p_0) \approx \sqrt{2\pi} \sqrt{n} \exp\left(-\frac{1}{2}t^2\right)$$

- ▶ **Key:** Bayes test will have the factor  $\sqrt{n}$   
This will asymptotically favour the null.
- ▶ There is only a big problem when  $2 < t < 4$  – **but** this is typically the most interesting case!

# Coin Tossing

**Intuition:** Imagine a coin tossing experiment and you want to determine whether the coin is “fair”  $H_0 : p = \frac{1}{2}$ .

There are four experiments.

Expt	1	2	3	4
n	50	100	400	10,000
r	32	60	220	5098
$L(p_0)$	0.81	1.09	2.17	11.68

# Coin Tossing

## Implications:

- ▶ Classical: In each case the  $t$ -ratio is approx 2. They we just  $H_0$  ( a fair coin) at the 5% level in each experiment.
- ▶ Bayes:  $L(p_0)$  grows to infinity and so they is overwhelming evidence for  $H_0$ . Connelly shows that the Monday effect disappears when you compute the Bayes version.

# Probability: 41901

## Week 5: Modern Regression Methods

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Topics

- ▶ Keynes vs Buffett: Regression
- ▶ Superbowl Spread Data
- ▶ Logistic Regression
- ▶ Least Squares
- ▶ Ridge Regression
- ▶ Lasso Regression

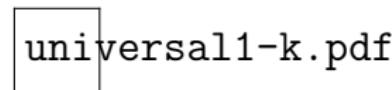
# Keynes versus Buffett

CAPM

$$\text{keynes} = 15.08 + 1.83 \text{ market}$$

$$\text{buffett} = 18.06 + 0.486 \text{ market}$$

Year	Keynes	Market
1928	-3.4	7.9
1929	0.8	6.6
1930	-32.4	-20.3
1931	-24.6	-25.0
1932	44.8	-5.8
1933	35.1	21.5
1934	33.1	-0.7
1935	44.3	5.3
1936	56.0	10.2
1937	8.5	-0.5
1938	-40.1	-16.1
1939	12.9	-7.2
1940	-15.6	-12.9
1941	33.5	12.5
1942	-0.9	0.8
1943	53.9	15.6
1944	14.5	5.4
1945	14.6	0.8



Keynes vs Cash

King's College Cambridge

# Shrinkage

## Basic Selection Principles

The goal is **model selection**

- ▶ Why not include all the variables?  
Big models tend to over-fit and find specific features.
- ▶ Need to trade-off fit for making good predictions.  
Friedman: A good model is one that predicts!

# MSE: Out-of-Sample Prediction

A very popular statistical criterion is  
mean squared error

MSE is defined by

$$MSE = \sum_Y (Y - \hat{Y})^2$$

- ▶ You make a prediction  $\hat{Y}$  about the variable  $Y$ .
- ▶ After the outcome  $Y$ , you calculate  $(Y - \hat{Y})^2$ .
- ▶ In data mining, it is popular to use a holdout sample and after you've built your statistical model to test it out-of-sample in terms of its mean squared error performance.

# Cross-Validation

Cross-Validation: Fit the model on training data.

Use model to predict  $\hat{Y}$ -values for the holdout sample.

Calculate predicted MSE  $\frac{1}{N} \sum_{j=1}^N (Y_j - \hat{Y}_j)^2$ .

Smallest MSE wins.

# Ridge Regression

Ridge Regression is a modification of the least squares criteria that minimizes (as a function of  $\beta$ 's):

$$\sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

for some value of  $\lambda > 0$

- ▶ The “blue” part of the equation is the traditional objective function of LS
- ▶ The “red” part is the shrinkage penalty, ie, something that makes costly to have big values for  $\beta$

# Ridge Regression

$$\sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p \beta_j^2$$

- ▶ if  $\lambda = 0$  we are back to least squares
- ▶ when  $\lambda \rightarrow \infty$ , it is “**too expensive**” to allow for any  $\beta$  to be different than 0...
- ▶ So, for different values of  $\lambda$  we get a different solution to the problem

# Ridge Regression

Ridge regression explores the **bias-variance trade-off!** The larger the  $\lambda$  the more bias (towards zero) is being introduced in the solution, ie, the less flexible the model becomes... at the same time, the solution has less **variance**

- ▶ As always, the trick to find the “right” value of  $\lambda$  that makes the model **not too simple but not too complex!**
- ▶ Whenever possible, we will choose  $\lambda$  by comparing the out-of-sample performance (usually via cross-validation)

# Ridge Regression

mse.png

# LASSO

The LASSO is a shrinkage method that performs automatic selection. Similar to ridge but it will provide solutions that are **sparse**, ie, some  $\beta$ 's exactly equal to 0! This facilitates interpretation of the results...

compare.png

# LASSO

The LASSO solves the following problem:

$$\arg \min \beta \left\{ \sum_{i=1}^n \left( Y_i - \beta_0 - \sum_{j=1}^p \beta_j X_{ij} \right)^2 + \lambda \sum_{j=1}^p |\beta_j| \right\}$$

- ▶ Once again,  $\lambda$  controls how flexible the model gets to be
- ▶ Still a very efficient computational strategy
- ▶ Whenever possible, we will choose  $\lambda$  by comparing the out-of-sample performance (usually via cross-validation).

# Ridge vs. LASSO

Why does the LASSO outputs zeros?

LASSO.png

# Ridge vs. LASSO

Which one is better?

- ▶ LASSO will perform better than Ridge when a relative small number of predictors have a strong effect in  $Y$  while Ridge will do better when  $Y$  is a function of many of the  $X$ 's and the coefficients are of moderate size
- ▶ LASSO can be easier to interpret (the zeros help!)
- ▶ But, if prediction is what we care about the only way to decide which method is better is comparing their out-of-sample performance

# Choosing $\lambda$

The idea is to solve the ridge or LASSO objective function over a grid of possible values for  $\lambda$ ...

lambda.png

# Optimization: Least Squares $L^2$ -norm

Gauss invented the concept of least squares

The  $L^2$ -regression objective function

$$\arg \min \beta ||y - X\beta||^2$$

Here parameter vector is  $\beta = (\beta_1, \dots, \beta_p)$ .

This has solution given by

$$\hat{\beta} = (X^\top X)^{-1} X^\top y$$

This can be numerically unstable when  $X^\top X$  is ill-conditioned.

Happens when  $p$  is large.

# Optimization: Ridge Regression

## $L^2 + L^2$ -norm

Ridge Regression has the solution

$$\hat{\beta}_{\text{ridge}} = (X^\top X + \lambda I)^{-1} X^\top y$$

You can plot the coefficients over a regularisation path of  $\lambda$ 's.

This can also be interpreted as a Bayesian hierarchical model with a normal likelihood and prior

# Optimization: LASSO $L^1$ -norm

## Least Absolute Shrinkage and Selection Operator (LASSO)

The solution to the lasso objective function

$$\arg \min \beta \left\{ \frac{1}{2} (y - \beta)^2 + \lambda |\beta| \right\}$$

is the soft-thresholding operator defined by

$$\hat{\beta} = \text{soft}(y; \lambda) = (y - \lambda \text{sgn}(y))_+$$

Here  $\text{sgn}$  is the sign function and  $(x)_+ = \max(x, 0)$ .

Define a slack variable  $z = |\beta|$  and solve the joint constrained optimisation problem which is differentiable

# Linear Regression

Model  $y_i = x_i^\top \beta + \epsilon_i$  where  $\beta = (\beta_1, \dots, \beta_p)$  for  $1 \leq i \leq n$ .

Equivalently:  $y = X\beta + \epsilon$  and  $\min \beta \|y - X\beta\|^2$ .

- ▶ Predictive Ability

The MLE (maximum likelihood) or OLS (Ordinary Least Squares) estimator is designed to have zero bias.

That means it can suffer with high variance!

There's a variance/bias tradeoff that we have to trade-off.

Main Advantage: Interpretability

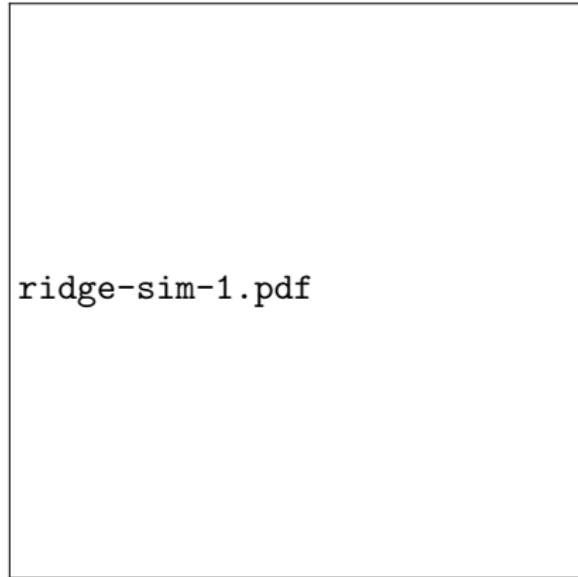
# Example

Simulated Data:  $n = 50, p = 30$  and  $\sigma^2 = 1$ .

True model: linear with 10 large coefficients between 0.5 and 1.

- ▶ Linear Regression
  - Bias squared = 0.006 and variance = 0.627.
  - Prediction error =  $1 + 0.006 + 0.627 = 1.633$
- ▶ We'll do better by shrinking the coefficients to reduce the variance
- ▶ How big a gain will we get with Ridge/Bayes Regression?

# Example: True Coefficients



ridge-sim-1.pdf

Figure: Shrinkage will Help

# Example: Prediction error

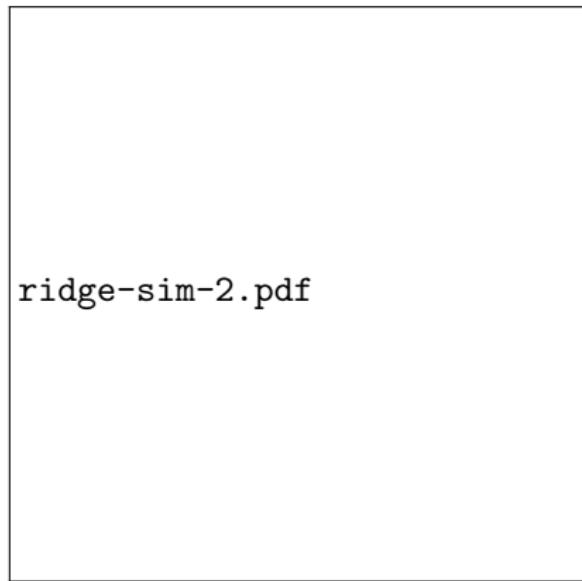


Figure: Ridge

Ridge Regression At best: Bias squared = 0.077 and variance = 0.402.  
Prediction error =  $1 + 0.077 + 0.403 = 1.48$

# Bias-Variance Tradeoff

Simulated data:  $n = 50, p = 30$ .

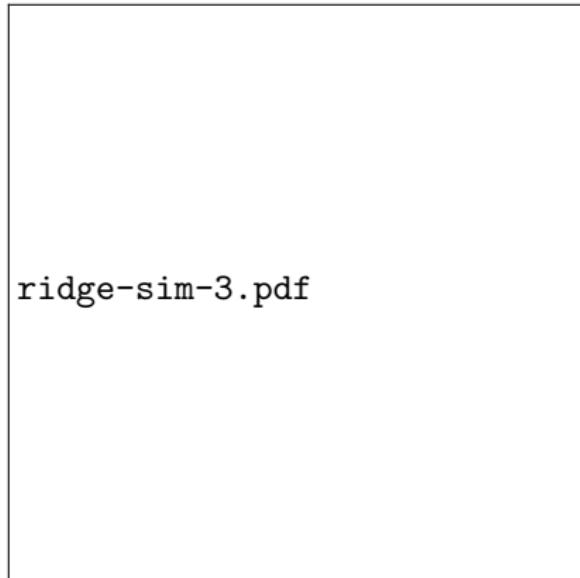


Figure: Ridge

# Example: Prostate Data and Regularisation Path

Cross-Validation: CV

prostate-lasso-plots.pdf

# What about Ridge?

prostate-ridge-plots.pdf

# Prostate Data

## CV Regularisation Path

Hastie, Friedman and Tibshirani (2013). *Elements of Statistical Learning*

prostate-lasso-plots.pdf

# What about Ridge?

prostate-ridge-plots.pdf

# Probability: 41901

## **Week 6: Bayesian Hierarchical Models and Finance**

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Hierarchical Models

Let's do some cool applications ...

- ▶ Bayes MoneyBall  
Batter-pitcher match-up:  
Kenny Lofton and Derek Jeter
- ▶ Bayes Elections
- ▶ SAT scores

# Example: Baseball

Batter-pitcher match-up?

Prior information on overall ability of a player.

Small sample size, pitcher variation.

- ▶ Let  $p_i$  denote Jeter's ability. Observed number of hits  $y_i$

$$(y_i|p_i) \sim Bin(T_i, p_i) \text{ with } p_i \sim Be(\alpha, \beta)$$

where  $T_i$  is the number of at-bats against pitcher  $i$ . A priori  $E(p_i) = \alpha / (\alpha + \beta) = \bar{p}_i$ .

- ▶ The extra heterogeneity leads to a prior variance  $Var(p_i) = \bar{p}_i(1 - \bar{p}_i)\phi$  where  $\phi = (\alpha + \beta + 1)^{-1}$ .

# Sports Data: Baseball

Table: Kenny Lofton hitting

Pitcher	At-bats	Hits	ObsAvg
J.C. Romero	9	6	.667
S. Lewis	5	3	.600
B. Tomko	20	11	.550
T. Hoffman	6	3	.500
K. Tapani	45	22	.489
A. Cook	9	4	.444
J. Abbott	34	14	.412
A.J. Burnett	15	6	.400
K. Rogers	43	17	.395
A. Harang	6	2	.333
K. Appier	49	15	.306
R. Clemens	62	14	.226
C. Zambrano	9	2	.222
N. Ryan	10	2	.200
E. Hanson	41	7	.171
E. Milton	19	1	.056
M. Prior	7	0	.000
Total	7630	2283	.299

Kenny Lofton versus individual pitchers.

# Baseball

Kenny Lofton

Kenny Lofton (career .299 average, and current .308 average for 2006 season) was facing the pitcher Milton (current record 1 for 19)

He was *rested* and replaced by a .275 hitter.

- ▶ Is putting in a weaker player really a better bet?
- ▶ Over-reaction to bad luck?

$$\mathbb{P} (\leq 1 \text{ hit in } 19 \text{ attempts} | p = 0.3) = 0.01$$

An unlikely 1-in-100 event.

# Baseball

Kenny Lofton

Bayes solution: shrinkage. Borrow strength across pitchers

Bayes estimate: use the posterior mean

Lofton's batting estimates that vary from .265 to .340.  
The lowest being against Milton.

.265 < .275

**Conclusion:** resting Lofton against Milton was justified!!

# Bayes Batter-pitcher match-up

Here's our model again ...

- ▶ Small sample sizes and pitcher variation.
- ▶ Let  $p_i$  denote Lofton's ability. Observed number of hits  $y_i$

$$(y_i|p_i) \sim Bin(T_i, p_i) \text{ with } p_i \sim Be(\alpha, \beta)$$

where  $T_i$  is the number of at-bats against pitcher  $i$ .

Estimate  $(\alpha, \beta)$

# Example: Derek Jeter

Table: Derek Jeter hierarchical model estimates

Pitcher	At-bats	Hits	ObsAvg	EstAvg	95% Int
R. Mendoza	6	5	.833	.322	(.282,.394)
H. Nomo	20	12	.600	.326	(.289,.407)
A.J.Burnett	5	3	.600	.320	(.275,.381)
E. Milton	28	14	.500	.324	(.291,.397)
D. Cone	8	4	.500	.320	(.218,.381)
R. Lopez	45	21	.467	.326	(.291,.401)
K. Escobar	39	16	.410	.322	(.281,.386)
J. Wettland	5	2	.400	.318	(.275,.375)
T. Wakefield	81	26	.321	.318	(.279,.364)
P. Martinez	83	21	.253	.312	(.254,.347)
K. Benson	8	2	.250	.317	(.264,.368)
T. Hudson	24	6	.250	.315	(.260,.362)
J. Smoltz	5	1	.200	.314	(.253,.355)
F. Garcia	25	5	.200	.314	(.253,.355)
B. Radke	41	8	.195	.311	(.247,.347)
D. Kolb	5	0	.000	.316	(.258,.363)
J. Julio	13	0	.000	.312	(.243,.350 )
Total	6530	2061	.316		

Derek Jeter 2006 season versus individual pitchers.

# Bayes Estimates

Stern estimates  $\hat{\phi} = (\alpha + \beta + 1)^{-1} = 0.002$  for Jeter

Doesn't vary much across the population of pitchers.

The extremes are shrunk the most also matchups with the smallest sample sizes.

Jeter had a season .308 average.

Bayes estimates vary from .311 to .327—he's very consistent.

If all players had a similar record then a constant batting average would make sense.

# Bayes Elections: Nate Silver

## Multinomial-Dirichlet

### Predicting the Electoral Vote (EV)

- ▶ Multinomial-Dirichlet:  $(\hat{p}|p) \sim Multi(p), (p|\alpha) \sim Dir(\alpha)$

$$p_{Obama} = (p_1, \dots, p_{51} | \hat{p}) \sim Dir(\alpha + \hat{p})$$

- ▶ Flat uninformative prior  $\alpha \equiv 1$ .

<http://www.electoral-vote.com/evp2012/Pres/prespolls.csv>

# Bayes Elections: Nate Silver

## Simulation

Calculate probabilities via simulation: `rdirichlet`

$$p(p_{j,O}|\text{data}) \text{ and } p(EV > 270|\text{data})$$

The election vote prediction is given by the sum

$$EV = \sum_{j=1}^{51} EV(j) \mathbb{E}(p_j|\text{data})$$

where  $EV(j)$  are for individual states

# Polling Data: electoral-vote.com

## Electoral Vote (EV), Polling Data: Mitt and Obama percentages

	State	M.pct	O.pct	EV				
1	Alabama	58	36	9	26	Missouri	48	48 11
2	Alaska	55	37	3	27	Montana	49	46 3
3	Arizona	50	46	10	28	Nebraska	60	34 5
4	Arkansas	51	44	6	29	Nevada	43	47 5
5	California	33	55	55	30	New.Hampshire	42	53 4
6	Colorado	45	52	9	31	New.Jersey	34	55 15
7	Connecticut	31	56	7	32	New.Mexico	43	51 5
8	Delaware	38	56	3	33	New.York	31	62 31
9	D.C.	13	82	3	34	North.Carolina	49	46 15
10	Florida	46	50	27	35	North.Dakota	43	45 3
11	Georgia	52	47	15	36	Ohio	47	45 20
12	Hawaii	32	63	4	37	Oklahoma	61	34 7
13	Idaho	68	26	4	38	Oregon	34	48 7
14	Illinois	35	59	21	39	Pennsylvania	46	52 21
15	Indiana	48	48	11	40	Rhode.Island	31	45 4
16	Iowa	37	54	7	41	South.Carolina	59	39 8
17	Kansas	63	31	6	42	South.Dakota	48	41 3
18	Kentucky	51	42	8	43	Tennessee	55	39 11
19	Louisiana	50	43	9	44	Texas	57	38 34
20	Maine	35	56	4	45	Utah	55	32 5
21	Maryland	39	54	10	46	Vermont	36	57 3
22	Massachusetts	34	53	12	47	Virginia	44	47 13
23	Michigan	37	53	17	48	Washington	39	51 11
24	Minnesota	42	53	10	49	West.Virginia	53	44 5
25	Mississippi	46	33	6	50	Wisconsin	42	53 10
	...					Wyoming	58	32 3

bayes-obama-2008.pdf

**Figure:** Election 2008 Prediction. Obama 370

bayes-obama-2012.pdf

**Figure:** Election 2012 Prediction. Obama 332.

# SAT Scores

SAT (200 – 800): 8 high schools and estimate effects.

School	Estimated $y_j$	St. Error $\sigma_j$	Average Treatment $\theta_i$
A	28	15	?
B	8	10	?
C	-3	16	?
D	7	11	?
E	-1	9	?
F	1	11	?
G	18	10	?
H	12	18	?

- ▶  $\theta_j$  average effects of coaching programs
- ▶  $y_j$  estimated treatment effects, for school  $j$ , standard error  $\sigma_j$ .

# Estimates

Two programs appear to work (improvements of 18 and 28)

- ▶ Large standard errors. Overlapping Confidence Intervals?
- ▶ Classical hypothesis test fails to reject the hypothesis that the  $\theta_j$ 's are equal.
- ▶ Pooled estimate has standard error of 4.2 with

$$\hat{\theta} = \frac{\sum_j (y_j / \sigma_j^2)}{\sum_j (1 / \sigma_j^2)} = 7.9$$

- ▶ Neither separate or pooled seems sensible.  
Bayesian shrinkage!

# Bayesian Model

Hierarchical Model ( $\sigma_j^2$  known) is given by

$$\bar{y}_j | \theta_j \sim N(\theta_j, \sigma_j^2)$$

Unequal variances–differential shrinkage.

- ▶ Prior Distribution:  $\theta_j \sim N(\mu, \tau^2)$  for  $1 \leq j \leq 8$ .  
Traditional random effects model.  
Exchangeable prior for the treatment effects.  
As  $\tau \rightarrow 0$  (complete pooling) and as  $\tau \rightarrow \infty$  (separate estimates).
- ▶ Hyper-prior Distribution:  $p(\mu, \tau^2) \propto 1/\tau$ .  
The posterior  $p(\mu, \tau^2|y)$  can be used to “estimate”  $(\mu, \tau^2)$ .

# Posterior

Joint Posterior Distribution  $y = (y_1, \dots, y_J)$

$$p(\theta, \mu, \tau | y) \propto p(y|\theta)p(\theta|\mu, \tau)p(\mu, \tau)$$

$$\propto p(\mu, \tau^2) \prod_{i=1}^8 N(\theta_j | \mu, \tau^2) \prod_{j=1}^8 N(y_j | \theta_j)$$

$$\propto \tau^{-9} \exp \left( -\frac{1}{2} \sum_j \frac{1}{\tau^2} (\theta_j - \mu)^2 - \frac{1}{2} \sum_j \frac{1}{\sigma_j^2} (y_j - \theta_j)^2 \right)$$

MCMC!

# Posterior Inference

Report posterior quantiles

School	2.5%	25%	50%	75%	97.5%
A	-2	6	10	16	32
B	-5	4	8	12	20
C	-12	3	7	11	22
D	-6	4	8	12	21
E	-10	2	6	10	19
F	-9	2	6	10	19
G	-1	6	10	15	27
H	-7	4	8	13	23
$\mu$	-2	5	8	11	18
$\tau$	0.3	2.3	5.1	8.8	21

Schools A and G are similar!

# Bayesian Shrinkage

Bayesian shrinkage provides a way of modeling complex datasets.

1. Coin Tossing: Lindley's Paradox
2. Baseball batting averages: Stein's Paradox
3. Toxoplasmosis
4. SAT scores
5. Clinical Trials

# Posterior

$p(p|\bar{y})$  is the posterior distribution for  $p$

$\bar{y}$  is a sufficient statistic.

- ▶ Bayes theorem gives

$$\begin{aligned} p(p|y) &\propto f(y|p)p(p|\alpha, \beta) \\ &\propto p^{\sum y_i} (1-p)^{n-\sum y_i} \cdot p^{\alpha-1} (1-p)^{\beta-1} \\ &\propto p^{\alpha+\sum y_i-1} (1-p)^{n-\sum y_i+\beta-1} \\ &\sim Beta(\alpha + \sum y_i, \beta + n - \sum y_i) \end{aligned}$$

- ▶ The posterior mean is a shrinkage estimator  
Combination of sample mean  $\bar{y}$  and prior mean  $E(p)$

$$E(p|y) = \frac{\alpha + \sum_{i=1}^n y_i}{\alpha + \beta + n} = \frac{n}{n + \alpha + \beta} \bar{y} + \frac{\alpha + \beta}{\alpha + \beta + n} \frac{\alpha}{\alpha + \beta}$$

# Poisson-Gamma

## Example

*Poisson/Gamma:* Suppose that  $Y_1, \dots, Y_n \sim Poi(\lambda)$ .

Let  $\lambda \sim Gamma(\alpha, \beta)$

$(\alpha, \beta)$  are known hyper-parameters.

- The **posterior** distribution is

$$\begin{aligned} p(\lambda|y) &\propto \exp(-n\lambda)\lambda^{\sum y_i}\lambda^{\alpha-1}\exp(-\beta\lambda) \\ &\sim Gamma(\alpha + \sum y_i, n + \beta) \end{aligned}$$

# Example: Clinical Trials

Novick and Grizzle: Bayesian Analysis of Clinical Trials

Four treatments for duodenal ulcers.

Doctors assess the state of the patient.

Sequential data

( $\alpha$ -spending function, can only look at prespecified times).

Treat	Excellent	Fair	Death
A	76	17	7
B	89	10	1
C	86	13	1
D	88	9	3

**Conclusion:** Cannot reject at the 5% level

Conjugate binomial/beta model+sensitivity analysis.

# Binomial-Beta

Let  $p_i$  be the death rate proportion under treatment  $i$ .

- ▶ To compare treatment  $A$  to  $B$  directly compute  $P(p_1 > p_2 | D)$ .
- ▶ Prior  $\text{beta}(\alpha, \beta)$  with prior mean  $E(p) = \frac{\alpha}{\alpha + \beta}$ .  
Posterior  $\text{beta}(\alpha + \sum x_i, \beta + n - \sum x_i)$
- ▶ For  $A$ ,  $\text{beta}(1, 1) \rightarrow \text{beta}(8, 94)$   
For  $B$ ,  $\text{beta}(1, 1) \rightarrow \text{beta}(2, 100)$
- ▶ Inference:  $P(p_1 > p_2 | D) \approx 0.98$

# Sensitivity Analysis

Important to do a sensitivity analysis.

Treat	Excellent	Fair	Death
A	76	17	7
B	89	10	1
C	86	13	1
D	88	9	3

Poisson-Gamma, prior  $\Gamma(m, z)$  and  $\lambda_i$  be the expected death rate.

Compute  $P\left(\frac{\lambda_1}{\lambda_2} > c | D\right)$

Prob	(0, 0)	(100, 2)	(200, 5)
$P\left(\frac{\lambda_1}{\lambda_2} > 1.3   D\right)$	0.95	0.88	0.79
$P\left(\frac{\lambda_1}{\lambda_2} > 1.6   D\right)$	0.91	0.80	0.64

# Normal-Normal Model

Using Bayes rule we get

$$p(\mu|y) \propto p(y|\mu)p(\mu)$$

- Posterior is given by

$$p(\mu|y) \propto \exp \left( -\frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mu)^2 - \frac{1}{2\tau^2} (\mu - \mu_0)^2 \right)$$

Hence  $\mu|y \sim N(\hat{\mu}_B, V_\mu)$  where

$$\hat{\mu}_B = \frac{n/\sigma^2}{n/\sigma^2 + 1/\tau^2} \bar{y} + \frac{1/\tau^2}{n/\sigma^2 + 1/\tau^2} \mu_0 \text{ and } V_\mu^{-1} = \frac{n}{\sigma^2} + \frac{1}{\tau^2}$$

A shrinkage estimator.

# Chicago Bears 2014-2015 Season

## Hierarchical Model

**Bayes Learning:** Update our beliefs in light of **new information**

- ▶ In the 2014-2015 season.  
The Bears suffered back-to-back 50-points defeats.  
Patriots-Bears 51 – 23  
Packers-Bears 55 – 14
- ▶ Their next game was at home against the Minnesota Vikings.  
Current line against the Vikings was **-3.5 points**.  
Slightly over a field goal

*What's the Bayes approach to learning the line?*

# Hierarchical Model

Hierarchical model for the current average win/lose this year

$$\begin{aligned}\bar{y}|\theta &\sim N\left(\theta, \frac{\sigma^2}{n}\right) \sim N\left(\theta, \frac{18.34^2}{9}\right) \\ \theta &\sim N(0, \tau^2)\end{aligned}$$

Here  $n = 9$  games so far. With  $s = 18.34$  points

Pre-season prior mean  $\mu_0 = 0$ , standard deviation  $\tau = 4$ .

Record so-far. Data  $\bar{y} = -9.22$ .

# Chicago Bears 2014-2015 Season

Bayes Shrinkage estimator

$$\mathbb{E}(\theta|\bar{y}, \tau) = \frac{\tau^2}{\tau^2 + \frac{\sigma^2}{n}} \bar{y}$$

The **Shrinkage factor** is 0.3!!

That's quite a bit of shrinkage. **Why?**

- ▶ Our updated estimator is

$$\mathbb{E}(\theta|\bar{y}, \tau) = -2.75 > -3.5$$

where current line is  $-3.5$ .

- ▶ Based on our hierarchical model this is an **over-reaction**.  
One point change on the line is about 3% on a probability scale.  
Alternatively, calculate a **market-based  $\tau$**  given line =  $-3.5$ .

# Chicago Bears

Last two defeats were 50 points scored by opponent

```
bears=c(-3,8,8,-21,-7,14,-13,-28,-41)
> mean(bears)
[1] -9.222222
> sd(bears)
[1] 18.34242
> tau=4

> sig2=sd(bears)*sd(bears)/9
> tau^2/(sig2+tau^2)
[1] 0.2997225
> 0.29997*-9.22
[1] -2.765723
> pnorm(-2.76/18)
[1] 0.4390677
```

Home advantage is worth 3 points. Vikings an average record.

**Result: Bears 21, Vikings 13**

# Shrinkage Estimation

**Stein paradox:** possible to make a uniform improvement on the MLE in terms of MSE.

- ▶ Mistrust of the statistical interpretation of Stein's result.  
In particular, the loss function.
- ▶ Difficulties in adapting the procedure to special cases
- ▶ Long familiarity with good properties for the MLE

Any gains from a “complicated” procedure could not be worth the extra trouble (Tukey, savings not more than 10 % in practice)

# Baseball Batting Averages

Data: 18 major-league players after 45 at bats (1970 season)

Player	$\bar{y}_i$	$E(p_i D)$	average season
Clemente	0.400	0.290	0.346
Robinson	0.378	0.286	0.298
Howard	0.356	0.281	0.276
Johnstone	0.333	0.277	0.222
Berry	0.311	0.273	0.273
Spencer	0.311	0.273	0.270
Kessinger	0.311	0.268	0.263
Alvarado	0.267	0.264	0.210
Santo	0.244	0.259	0.269
Swoboda	0.244	0.259	0.230
Unser	0.222	0.254	0.264
Williams	0.222	0.254	0.256
Scott	0.222	0.254	0.303
Petrocelli	0.222	0.254	0.264
Rodriguez	0.222	0.254	0.226
Campanens	0.200	0.259	0.285
Munson	0.178	0.244	0.316
Alvis	0.156	0.239	0.200

# Baseball Data

First Shrinkage Estimator: Efron and Morris

`efron-morris-baseball.pdf`

# Shrinkage

Let  $\theta_i$  denote the end of season average

- ▶ Lindley: shrink to the overall grand mean

$$c = 1 - \frac{(k - 3)\sigma^2}{\sum(\bar{y}_i - \bar{y})^2}$$

where  $\bar{y}$  is the overall grand mean and

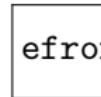
$$\hat{\theta} = c\bar{y}_i + (1 - c)\bar{y}$$

- ▶ Baseball data:  $c = 0.212$  and  $\bar{y} = 0.265$ .  
Compute  $\sum(\hat{\theta}_i - \bar{y}_i^{obs})^2$  and see which is lower:

$$MLE = 0.077 \text{ STEIN} = 0.022$$

That's a factor of 3.5 times better!

# Batting Averages



efron-clemente.pdf

# Baseball Paradoxes

Shrinkage on Clemente too severe:

$$z_{CI} = 0.265 + 0.212(0.400 - 0.265) = 0.294.$$

The 0.212 seems a little severe

- ▶ Limited translation rules, maximum shrinkage eg. 80%
- ▶ Not enough shrinkage eg O'Connor ( $y = 1, n = 2$ ).  
$$z_{O'C} = 0.265 + 0.212(0.5 - 0.265) = 0.421.$$
 Still better than Ted Williams 0.406 in 1941.
- ▶ Foreign car sales ( $k = 19$ ) will further improve MSE performance! It will change the shrinkage factors.
- ▶ Clearly an improvement over the Stein estimator is

$$\hat{\theta}_{S+} = \max \left( \left( 1 - \frac{k-2}{\sum \bar{Y}_i^2} \right), 0 \right) \bar{Y}_i$$

# Baseball Prior

Include extra prior knowledge

Empirical distribution of all major league players

$$\theta_i \sim N(0.270, 0.015)$$

The 0.270 provides another origin to shrink to and the prior variance 0.015 would give a different shrinkage factor.

To fully understand maybe we should build a probabilistic model and see what the posterior mean is as our estimator for the unknown parameters.

# Shrinkage: Unequal Variances

Model  $Y_i|\theta_i \sim N(\theta_i, D_i)$  where  $\theta_i \sim N(\theta_0, A) \sim N(0.270, 0.015)$ .

- ▶ The  $D_i$  can be different – unequal variances
- ▶ Bayes posterior means are given by

$$E(\theta_i|Y) = (1 - B_i)Y_i \text{ where } B_i = \frac{D_i}{D_i + A}$$

where  $\hat{A}$  is estimated from the data, see Efron and Morris (1975).

- ▶ Different shrinkage factors as different variances  $D_i$ .  
 $D_i \propto n_i^{-1}$  and so smaller sample sizes are shrunk more.  
Makes sense.

# Example: Toxoplasmosis Data

Disease of Blood that is endemic in tropical regions.

Data: 5000 people in El Salvador (varying sample sizes) from 36 cities.

- ▶ Estimate “true” prevalences  $\theta_i$  for  $1 \leq i \leq 36$
- ▶ Allocation of Resources: should we spend funds on the city with the highest observed occurrence of the disease? Same shrinkage factors?
- ▶ Shrinkage Procedure (Efron and Morris, p315)

$$z_i = c_i y_i$$

where  $y_i$  are the observed relative rates (normalized so  $\bar{y} = 0$ )  
The smaller sample sizes will get shrunk more.

The most gentle are in the range  $0.6 \rightarrow 0.9$  but some are  
 $0.1 \rightarrow 0.3$ .

# Bayes Portfolio Selection

de Finetti and Markowitz: Mean-variance portfolio shrinkage:

$$\frac{1}{\gamma} \Sigma^{-1} \mu$$

Different shrinkage factors for different history lengths.

Portfolio Allocation in the SP500 index

Entry/exit; splits; spin-offs etc. For example, 73 replacements to the SP500 index in period 1/1/94 to 12/31/96.

**Advantage:**  $E(\alpha|D_t) = 0.39$ , that is 39 bps per month which on an annual basis is  $\alpha = 468$  bps.

The posterior mean for  $\beta$  is  $p(\beta|D_t) = 0.745$

$\bar{x}_M = 12.25\%$  and  $\bar{x}_{PT} = 14.05\%$ .

# SP Composition

Date	Symbol	6/96	12/89	12/79	12/69
General Electric	GE	2.800	2.485	1.640	1.569
Coca Cola	KO	2.342	1.126	0.606	1.051
Exxon	XON	2.142	2.672	3.439	2.957
ATT	T	2.030	2.090	5.197	5.948
Philip Morris	MO	1.678	1.649	0.637	*****
Royal Dutch	RD	1.636	1.774	1.191	*****
Merck	MRK	1.615	1.308	0.773	0.906
Microsoft	MSFT	1.436	*****	*****	*****
Johnson/Johnson	JNJ	1.320	0.845	0.689	*****
Intel	INTC	1.262	*****	*****	*****
Procter and Gamble	PG	1.228	1.040	0.871	0.993
Walmart	WMT	1.208	1.084	*****	*****
IBM	IBM	1.181	2.327	5.341	9.231
Hewlett Packard	HWP	1.105	0.477	0.497	*****
Pepsi	PEP	1.061	0.719	*****	*****
Pfizer	PFE	0.918	0.491	0.408	0.486
Dupont	DD	0.910	1.229	0.837	1.101
AIG	AIG	0.910	0.723	*****	*****
Mobil	MOB	0.906	1.093	1.659	1.040
Bristol Myers Squibb	BMY	0.878	1.247	*****	0.484
GTE	GTE	0.849	0.975	0.593	0.705
General Motors	GM	0.848	1.086	2.079	4.399
Disney	DIS	0.839	0.644	*****	*****
Citicorp	CCI	0.831	0.400	0.418	*****
BellSouth	BLS	0.822	1.190	*****	*****
Motorola	MOT	0.804	*****	*****	*****

# SP Composition

Ford	F	0.798	0.883	0.485	0.640
Chervon	CHV	0.794	0.990	1.370	0.966
Amoco	AN	0.733	1.198	1.673	0.758
Eli Lilly	LLY	0.720	0.814	*****	*****
Abbott Labs	ABT	0.690	0.654	*****	*****
AmerHome Products	AHP	0.686	0.716	0.606	0.793
FedNatlMortgage	FNM	0.686	*****	*****	*****
McDonald's	MCD	0.686	0.545	*****	*****
Ameritech	AIT	0.639	0.782	*****	*****
Cisco Systems	CSCO	0.633	*****	*****	*****
CMB	CMB	0.621	*****	*****	*****
SBC	SBC	0.612	0.819	*****	*****
Boeing	BA	0.598	0.584	0.462	*****
MMM	MMM	0.581	0.762	0.838	1.331
BankAmerica	BAC	0.560	*****	0.577	*****
Bell Atlantic	BEL	0.556	0.946	*****	*****
Gillette	G	0.535	*****	*****	*****
Kodak	EK	0.524	0.570	1.106	*****
Chrysler	C	0.507	*****	*****	0.367
Home Depot	HD	0.497	*****	*****	*****
Colgate	COL	0.489	0.499	*****	*****
Wells Fargo	WFC	0.478	*****	*****	*****
Nations Bank	NB	0.453	*****	*****	*****
Amer Express	AXP	0.450	0.621	*****	*****

# Probability: 41901

## Week 8: Sports Betting and Brownian Motion

Nick Polson

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# Overview

Kelly-Breiman-Merton: **Maximise Expected Growth Rate**

1. Sir Isaac Newton: Lost £20,000.

*"I can calculate the motion of heavenly bodies, but not the madness of people"* South Sea Bubble Act (1721)

Newton Master of the Mint (1700-1727). Gold standard.

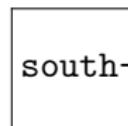
2. David Ricardo (1815)

*Bought all issuance of UK Gilts the day before the Battle of Waterloo*

3. J.M. Keynes (1920-1945)

*King's College Fund.*

# South Sea Bubble: 1720

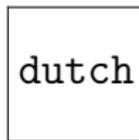


south-sea-newton.jpg

*Also owned the East India Company: £30,000*

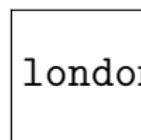
# Post 1720 Bubble

Volatility: persistent, asymmetric (leverage) and fat-tails



dutch.jpg

Dutch East India Company



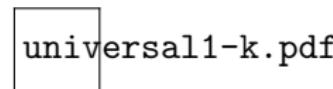
london.jpg

London Stock Market:  
1723-1794

# Keynes Investment Returns

1928-1945

Year	Keynes	Market
1928	-3.4	7.9
1929	0.8	6.6
1930	-32.4	-20.3
1931	-24.6	-25.0
1932	44.8	-5.8
1933	35.1	21.5
1934	33.1	-0.7
1935	44.3	5.3
1936	56.0	10.2
1937	8.5	-0.5
1938	-40.1	-16.1
1939	12.9	-7.2
1940	-15.6	-12.9
1941	33.5	12.5
1942	-0.9	0.8
1943	53.9	15.6
1944	14.5	5.4
1945	14.6	0.8



Keynes vs Cash

King's College Cambridge

# Keynes

## Optimal Bayes Rebalancing



universal2-k.pdf

Optimal allocation  $\omega$



universal3-k.pdf

Keynes vs Universal vs Cash

# Parrando's Paradoxes

Two losing bets can be combined to a winner

Bernoulli:  $1+f$  or  $1-f$  with  $p = 0.51$  and  $f = 0.05$

Caveat: Growth governed by entropy

$$p \log(1+f) + (1-p) \log(1-f) = -0.00025 < 0$$

Brownian Ratchets and cross-entropy of Markov processes

 universal1.pdf

Two Losing Bets+Volatility

# Parrando



universal2.pdf

Optimal allocation  $\omega$



universal3.pdf

Ex Ante vs Ex Post

# Breiman-Kelly-Merton Rule

Kelly Criterion: Optimal wager in binary setting

$$\omega^* = \frac{p \cdot O - q}{O}$$

Merton's Rule in continuous setting is Kelly

$$\omega^* = \frac{1}{\gamma} \frac{\mu}{\sigma^2}$$

1.  $\mu$ : (excess) expected return
2.  $\sigma$ : volatility
3.  $\gamma$ : risk aversion
4.  $\omega^*$ : optimal position size
5.  $p = \text{Prob}(Up), q = \text{Prob}(Down), O = \text{Odds}$

# Example: Kelly Criterion S&P500:

Kelly rule is logarithmic utility (CRRA with  $\gamma = 1$ ).

- ▶ Given i.i.d. log-normal stock returns with an annualized expected excess return of 5.7% and a volatility of 16% which is consistent with long-run equity returns.

$$\omega^* = 0.057 / 0.16^2 = 2.22$$

Kelly implies the investor borrows 122% of wealth to invest a total of 220% in stocks.

- ▶ Allocation is highly sensitive to estimation error in  $\hat{\mu}$ .  
Dynamic learning?

# Fractional Kelly

The fractional Kelly rule leads to a more realistic allocation.

- ▶ Suppose that  $\gamma = 3$ . Then the informational ratio is

$$\frac{\mu}{\sigma} = \frac{0.057}{0.16} = 0.357 \text{ and } \omega^* = \frac{1}{3} \frac{0.057}{0.16^2} = 74.2\%$$

An investor with such a level of risk aversion then has a more reasonable 74.2% allocation.

- ▶ This analysis ignores the equilibrium implications. If every investor acted this way, then this would drive up prices and drive down the equity premium of 5.7%.

# Black-Litterman

Black-Litterman: combining investor's Bayes views with market equilibrium

- ▶ Optimal allocation rule is

$$\omega^* = \frac{1}{\gamma} \Sigma^{-1} \mu$$

Q. How to specify  $(\mu, \Sigma)$  pairs?

- ▶ Given  $\hat{\Sigma}$ , BL derive Bayesian inference for  $\mu$  given market equilibrium model and *a priori* views on the returns of pre-specified portfolios:

$$(\hat{\mu}|\mu) \sim \mathcal{N}(\mu, \tau \hat{\Sigma}) \text{ and } (Q|\mu) \sim \mathcal{N}(P\mu, \hat{\Omega}) .$$

# Posterior Views

Combining views, the implied posterior is

$$(\mu | \hat{\mu}, Q) \sim \mathcal{N}(Bb, B)$$

The mean and variance are specified by

$$B = (\tau \hat{\Sigma})^{-1} + P' \hat{\Omega}^{-1} P \text{ and } b = (\tau \hat{\Sigma})^{-1} \hat{\mu} + P' \Omega^{-1} Q$$

These posterior moments then define the optimal allocation rule.

# Brownian Motion

Stochastic process  $\{B_t, t \geq 0\}$  in continuous time taking real values is a Brownian Motion (or a Wiener process)

- ▶ for each  $s \geq 0$  and  $t > 0$

$$B_{t+s} - B_t \sim N(0, t).$$

- ▶ For each  $t > s > 0$ , the random variable  $B_t - B_s$  is independent of  $B_s - B_0 = B_s$ .
- ▶  $B_0 = 0$  and
- ▶  $B_t$  is continuous in  $t \geq 0$ .
- ▶  $B_t$  is standard BM.  $\sigma B_t$  includes a volatility.

Brown (1827) and Einstein (1921).

# Brownian Motion with Drift

- Here we have the process

$$X_t = \mu t + \sigma B_t$$
$$dX_t = \mu dt + \sigma dB_t$$

- We can think of the Euler discretization as

$$X_{t+\Delta} - X_t = \mu\Delta + \sigma\sqrt{\Delta}\epsilon_t$$

where  $\epsilon_t \sim \mathcal{N}(0, 1)$ .

# Geometric Brownian Motion

- Geometric Brownian Motion starting at  $X_0 = 1$  evolves as

$$X_t = \exp(\mu t + \sigma B_t)$$

We have  $\mathbb{E}(X_t) = e^{(\mu + \frac{1}{2}\sigma^2)t}$ .

- This is equivalent differential (SDE) form

$$dX_t = X_t \left( \left( \mu + \frac{1}{2}\sigma^2 \right) dt + \sigma dB_t \right)$$

# Brownian Motion for Sports Scores

## Implied Volatility

Stern (1994): *Brownian Motion and Sports Scores*

Polson and Stern (2014): *The Implied Volatility of a Sports Game*

- ▶ “Team A” rarely loses if they are ahead at halftime
- ▶ Approximately 75% of games are won by the team that leads at halftime.

There's a 0.5 probability that the same team wins both halves and a 0.25 probability that the first half winner defeats the second half winner.

This only applies to equally matched teams.

- ▶ For the team that is ahead after  $\frac{3}{4}$  of the game, the empirical frequency of winnings is: basketball 80%, baseball 90%, football 80% and hockey 80%.

# Model Checking

$\mu$  home point advantage.

Football 3 points, basketball 5 – 6 points.

- ▶ Stern estimates for basketball a 1.5 points difference for the first three quarters with a standard deviation of 7.5 points.  
Fourth quarter the difference is only 0.22 points with a standard deviation of 7.

The q-q plot look normals. Correlations between periods are small and so the random walk model appears to be reasonable.

# Evolution of a Game

## Volatility of Outcome

hal-vol-new.pdf

# Brownian Motion for Sports Scores

Probability of team A winning,  $p = \mathbb{P}(X(1) > 0)$ , given point spread (or drift)  $\mu$

Standard deviation (or volatility)  $\sigma$  and final score  $X(1) \sim N(\mu, \sigma^2)$ .

- ▶ Given the normality assumption,  $X(1) \sim N(\mu, \sigma^2)$ , we have

$$p = \mathbb{P}(X(1) > 0) = \Phi\left(\frac{\mu}{\sigma}\right)$$

where  $\Phi$  is the standard normal cdf.

# Brownian Motion for Sports Scores

Table 1 uses  $\Phi$  to convert team A's advantage  $\mu$  to a probability scale using the information ratio  $\mu/\sigma$ .

$\mu/\sigma$	0	0.25	0.5	0.75	1	1.25	1.5	2
$p = \Phi(\mu/\sigma)$	0.5	0.60	0.69	0.77	0.84	0.89	0.93	0.977

Probability of winning  $p$  versus the ratio  $\mu/\sigma$

Evenly matched and  $\mu/\sigma = 0$  then  $p = 0.5$ .

# Conditional Probability

After time  $t$  has elapsed:

- ▶ Current lead of  $l$  for team A as the conditional distribution:

$$(X(1)|X(t) = l) = (X(1) - X(t)) + l \sim N(l + \mu(1 - t), \sigma^2(1 - t))$$

- ▶ The probability of team A winning at time  $t$  given a current lead of  $l$  points is:

$$p_t = P(X(1) > 0 | X(t) = l) = \Phi\left(\frac{l + \mu(1 - t)}{\sigma\sqrt{1 - t}}\right)$$

- ▶ Point spread  $\mu = -4$  and volatility is  $\sigma = 10.6$ , then team A has a  $\mu/\sigma = -4/10.6 = -0.38$  volatility point disadvantage.  
The probability of winning is  $\Phi(-0.38) = 0.353 < 0.5$ .

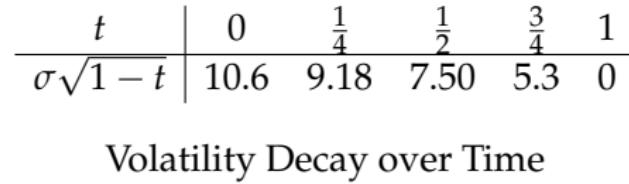
# Implied Volatility

Initial point spread–markets' expectation– of outcome.

Probability that the underdog team wins is

$$p = \Phi(\mu/\sigma) = \Phi(-4/10.6) = 35.3\%.$$

- ▶ The volatility is a decreasing function of  $t$ , illustrating that the volatility dissipates over the course of a game.



Calculate *implied volatility*,  $\sigma_{IV}$ , by solving

$$\sigma_{IV} : \quad p = \Phi\left(\frac{\mu}{\sigma_{IV}}\right) \text{ which gives } \sigma_{IV} = \frac{\mu}{\Phi^{-1}(p)} .$$

# SuperBowl XLVII: Ravens vs 49ers

TradeSports.com

hal-superbowl.png

# Super Bowl XLVII: Ravens vs 49ers

Super Bowl XLVII was held at the Superdome in New Orleans on February 3, 2013.

We will track  $X(t)$  which corresponds to the Raven's lead over the 49ers at each point in time. Table 3 provides the score at the end of each quarter.

$t$	0	$\frac{1}{4}$	$\frac{1}{2}$	$\frac{3}{4}$	1
Ravens	0	7	21	28	34
49ers	0	3	6	23	31
$X(t)$	0	4	15	5	3

SuperBowl XLVII by Quarter

# Initial Market

Initial *point spread* which was set at the Ravens being a four point underdog,  $\mu = -4$ .

$$\mu = \mathbb{E}(X(1)) = -4.$$

The Ravens upset the 49ers by  $34 - 31$  and  $X(1) = 34 - 31 = 3$  with the point spread being beaten by 7 points.

- ▶ To determine the markets' assessment of the probability that the Ravens would win at the beginning of the game we use the *money-line* odds.

San Francisco –175 and Baltimore Ravens +155. This implies that a bettor would have to place \$175 to win \$100 on the 49ers and a bet of \$100 on the Ravens would lead to a win of \$155.

Convert both of these money-lines to *implied probabilities* of the each team winning

$$p_{SF} = \frac{175}{100 + 175} = 0.686 \text{ and } p_{Bal} = \frac{100}{100 + 155} = 0.392$$

# Probabilities of Winning

The probabilities do not sum to one. This "excess" probability is in fact the mechanism by which the oddsmakers derive their compensation.

The difference is the "*market vig*", also known as the bookmaker's edge.

$$p_{SF} + p_{Bal} = 0.686 + 0.392 = 1.078$$

providing a 7.8% edge for the bookmakers.

- ▶ Put differently, if bettors place money proportionally across both teams then the bookies will make 7.8% of the total staked.  
We use the mid-point of the spread to determine  $p$  implying that

$$p = \frac{1}{2}p_{Bal} + \frac{1}{2}(1 - p_{SF}) = 0.353$$

From the Ravens perspective, we have

$$p = \mathbb{P}(X(1) > 0) = 0.353.$$

Baltimore's win probability started trading at  $p_0^{mkt} = 0.38$

# Half Time Analysis

The Ravens took a commanding 21 – 6 lead at half time.

- ▶ The market was trading at  $p_{\frac{1}{2}}^{mkt} = 0.90$ .

During the 34 minute blackout 42760 contracts changed hands with Baltimore's win probability ticking down from 95 to 94.

The win probability peak of 95% occurred again after a third-quarter kickoff return for a touchdown.

At the end of the four quarter, however, when the 49ers nearly went into the lead with a touchdown, Baltimore's win probability had dropped to 30%.

# Implied Volatility

To calculate the implied volatility of the Superbowl we substitute the pair  $(\mu, p) = (-4, .353)$  into our definition and solve for  $\sigma_{IV}$ .

$$\sigma = \frac{\mu}{\Phi^{-1}(p)},$$

- We obtain

$$\sigma_{IV} = \frac{\mu}{\Phi^{-1}(p)} = \frac{-4}{-0.377} = 10.60$$

where  $\Phi^{-1}(p) = qnorm(0.353) = -0.377$ . The 4 point advantage assessed for the 49ers is under a  $\frac{1}{2}\sigma$  favorite.

- The outcome  $X(1) = 3$  was within one standard deviation of the pregame model which had an expectation of  $\mu = -4$  and volatility of  $\sigma = 10.6$ .

# Half Time Probabilities

*What's the probability of the Ravens winning given their lead at half time?*  
At half time Baltimore led by 15 points, 21 to 6.

The conditional mean for the final outcome is  $15 + 0.5 * (-4) = 13$   
and the conditional volatility is  $10.6\sqrt{1-t}$ .

These imply a probability of .96 for Baltimore to win the game.

- ▶ A second estimate of the probability of winning given the half time lead can be obtained directly from the betting market.  
From the online betting market we also have traded contracts on TradeSports.com that yield a half time probability of  $p_{\frac{1}{2}} = 0.90$ .

# *What's the implied volatility for the second half?*

$p_t^{mkt}$  reflects all available information

- ▶ For example, at half-time  $t = \frac{1}{2}$  we would update

$$\sigma_{IV,t=\frac{1}{2}} = \frac{l + \mu(1-t)}{\Phi^{-1}(p_t)\sqrt{1-t}} = \frac{15 - 2}{\Phi^{-1}(0.9)/\sqrt{2}} = 14$$

where  $qnorm(0.9) = 1.28$ .

- ▶ As  $14 > 10.6$ , the market was expecting a more volatile second half—possibly anticipating a comeback from the 49ers.

# How can we form a valid betting strategy?

Given the initial implied volatility  $\sigma = 10.6$ .

At half time with the Ravens having a  $l + \mu(1 - t) = 13$  points edge

- ▶ We would assess with  $\sigma = 10.6$

$$p_{\frac{1}{2}} = \Phi \left( 13 / (10.6 / \sqrt{2}) \right) = 0.96$$

probability of winning versus the  $p_{\frac{1}{2}}^{mkt} = 0.90$  rate.

- ▶ To determine our optimal bet size,  $\omega_{bet}$ , on the Ravens we might appeal to the Kelly criterion (Kelly, 1956) which yields

$$\omega_{bet} = p_{\frac{1}{2}} - \frac{q_{\frac{1}{2}}}{O^{mkt}} = 0.96 - \frac{0.1}{1/9} = 0.60$$

bball-choosingsigma.pdf



brownian-fan-fig1.pdf

**Figure:** The family of probability distributions for  $X(t)$ , for any  $t$  between 0 and 1. Any vertical slice for a specific value  $t_1$  corresponds to a normal distribution for  $X(t_1)$ . The darker the shading at  $t$ , the higher the probability of the corresponding score-difference for  $X(t)$  along that particular time slice.

# Probability: 41901

## Week 8: AI

Nick Polson

The University of Chicago Booth School of Business

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

# AI: Introduction

What Does “AI” Really Mean?

Think of an algorithm.

Two distinguishing features of AI algorithms:

1. Algorithms typically deal with probabilities rather than certainties.
2. There's the question of how these algorithms “know” what instructions to follow.

# Three AI enablers

1. Moore's law:  
Decades-long exponential growth in the speed of computers  
(Intel, Nvidia)
2. New Moore's law:  
Explosive growth in the amount of data, as all of humanity's information is digitized
3. Cloud-computing (Nvidia, Google, AWS, Facebook, Azure)

# Three Factors Driving AI

## 1. Massive Data

Fitting complicated models to describe complicated patterns without overfitting requires millions or billions of data points.

## 2. Trial and Error.

A Billion Times per Second (Chess, Go). Better than 10,000 hours.

## 3. Deep Learning

Pattern Recognition

# AI: Pattern Recognition

How does Zillow predicts the price of a house?

How do beekeepers save honeybees?

How to sort cucumbers?

– Pattern Recognition

Good prediction rule to map input to output

# Examples of Image Classification

Pattern: relationship between the visual features and its class.

- ▶ Cucumber sorter of images to sort them into nine different classes.
- ▶ Toilet paper assignment at the Temple of Heaven in Beijing
- ▶ Identification of untagged friends on Facebook
- ▶ Detection collisions between subatomic particles at CERN

Ultimately, computers are agnostic about the type of input you give them, because to a computer, it's all just numbers.

# Examples of Pattern Recognition

Input	Output
02_soundwave.png	Speech to text: “Chi-ca-go hot-dog.”
68F/20C, 70% humidity, mostly sunny	Numerical prediction: “Power consumption in Lon- don will be 25,500 megawatt- hours.”

Google Assistant

# Two key ideas behind pattern-recognition systems

1. In AI, a “pattern” is a prediction rule mapping an input to expected output.
2. “Learning a pattern” means fitting a good prediction rule to a data set.

In AI, prediction rules are often referred to as “models”. The process of using data to find a good prediction rule is called “training the model”.

## Ethics of Automating your job

# Chatbot lawyer overturns 160,000 parking tickets in London and New York

DoNotPay has taken on 250,000 cases and won 160,000  
Giving a success rate of 64%

The bot was created by the self-taught coder after receiving 30 parking tickets at the age of 18 in and around London.

The process for appealing the fines is relatively formulaic and perfectly suits AI, which is able to quickly give the appropriate advice without charging lawyers fees

# Smart cities. Copenhagen bicycles.

Copenhagen wants to cut bus travel times by 5 to 20 percent and cycling travel times by 10 percent. Reduce the number of times cyclists have to stop by 10 percent.

To better manage the traffic, there's an AI that identifies cars/bicyclists and gives priority to cyclists at morning peak hours. Installing 380 "intelligent traffic signals" that will spot, and prioritize, buses and bikes.

# David Bowie and Verbasizer

David Bowie used machine learning (ML) to write his songs.  
Bowie comments that the questionable quality of the lyrics from his old band Tin Machine should be “blamed” on the computer.

Verbasizer automates a technique used by Bowie to write songs

Form random phrases based on shuffled word cut-outs from newspapers and other sources.

# Japanese Farming: Cucumbers

TensorFlow

Makoto Koike sorts cucumbers at his parents farm. 9 classes of cucumbers.

His mother spends up to eight hours per day at peak harvesting times.

Learning to sort cucumbers can take months. Size, thickness, color, texture, scratches, whether or not they are crooked or have prickles.

- ▶ Data 7000 pictures of cucumbers sorted by his mother. What are the important “features”?
- ▶ Recognition accuracy exceeded 95% for test images, out-of-sample the accuracy drops to about 70%.

Farming and TensorFlow

# Predictive Analytics

## General Introduction

Predictive Analytics is the most widely used tool for high dimensional input-output

$$Y = F(X) \text{ where } X = (X_1, \dots, X_p)$$

- ▶ Consumer Demand (Amazon, Airbnb, ... )
- ▶ Maps (Bing, Uber)
- ▶ Pricing
- ▶ Healthcare

The applications are endless ....

# Target

Target and other retailers use predictive analytics to study consumer purchasing behaviour to see what type of coupons or promotions you might like

Here's a famous story about a father and his daughter. Target predicted that his daughter was pregnant from her purchasing behaviour long before they were buying diapers

Here's the original link ...

Target and Pregnancy

# Walmart

Walmart began using predictive analytics in 2004. Mining trillions of bytes' worth of sales data from recent hurricanes

Determine what customers most want to purchase leading up to a storm.

Strawberry Pop-Tarts are one of the most purchased food items, especially after storms, as they require no heating and can be eaten at any meal

Walmart and Hurricanes

# Airbnb

## Airbnb New User Bookings Prediction Competition

New users on Airbnb can book a place to stay in 34,000+ cities across 190+ countries.

Accurately predict where a new user will book their first travel experience

Airbnb can then personalized content, decrease the average time to first booking, and better forecast demand.

12 classes—major destinations, and a did not book category

# Airbnb

List of users, demographics, web session records, and content data

`airbnb_comp.png`

Winner has the best out-of-sample prediction!!

# Germany's Otto

## Predicting Consumer Demand

Customers were less likely to return merchandise if it arrived within two days.

If an item takes longer to arrive, it gives customers more time to spot the product in a shop for less money and buy it, forcing Otto to forgo the sale and eat the shipping costs.

While customers are less likely to return merchandise that arrives quickly, also they prefer to receive everything at once.

# Personalized Recommendations for Experiences Using Deep Learning

As the number and kinds of available experiences on TripAdvisor grew rapidly in the last couple of years, a personalized website can increase user satisfaction significantly by providing travelers with an easy way to find experiences that are relevant to them.

- ▶ Approach
  1. Training Data Collection
  2. Entity Embeddings
  3. RFY Model Architecture
- ▶ Result Analysis
  1. Offline Evaluation
  2. Online A/B test

# Germany's Otto

Otto sells other brands, does not stock those goods itself, hard to avoid one of the two evils: shipping delays until all the orders are ready for fulfilment, or lots of boxes arriving at different times.

- ▶ Analyze around 3 billion past transactions and 200 variables—past sales, searches on Otto's site and weather information. They predict what customers will buy a week before they order. This system has proved so reliable, predicting with 90% accuracy what will be sold within 30 days, that Otto allows it automatically to purchase around 200,000 items a month from third-party brands with no human intervention.

Economist  
Germany's Otto

# Stitch Fix CEO Says AI Is 'So Real' and Incredibly Valuable

Stitch Fix asks customers for insights and feedback alongside their size and color preference for items, even the ones customers didn't like or buy, in exchange for a clear value proposition.

The breadth and depth of their data are valuable.

Their model relies on a combination of data science – machine learning, AI and natural language processing – and human stylists; on top of complex customer profiles built by data, stylists can layer the nuances of buying and wearing clothes.

# Uber Pool

Bayes predicts where you're going to be dropped off.

Uber constructs prior probabilities for riders, Uber cars, and popular places.

Combine to construct a joint probability table

Then calculate the posterior probability of destination for each person and pool travellers together

Uber Pool

# Probability: 41901

## Week 9 & 10: Deep Learning

Nick Polson

The University of Chicago Booth School of Business

<http://faculty.chicagobooth.edu/nicholas.polson/teaching/41901/>

Suggested Reading  
AWS/Nvidia/Google DeepMind

# Deep Learning: Introduction

Deep Learning is the most widely used machine learning tool for high dimensional input-output problems

- ▶ Image Recognition
- ▶ Google Translate
- ▶ Driverless Cars

The applications are endless ....

*If you can define the task, I can build a machine to do it*

*John von Neumann*

# Kolmogorov-Arnold

There are no multivariate functions just superpositions of univariate ones

Let  $f_1, \dots, f_L$  be given univariate activation functions. We set

$$f_l^{W,b} = f_l \left( \sum_{j=1}^{N_l} W_{lj} X_j + b_l \right) = f_l(W_l X_l + b_l), \quad 1 \leq l \leq L,$$

Deep predictor has hidden units  $N_l$  and depth  $L$ .

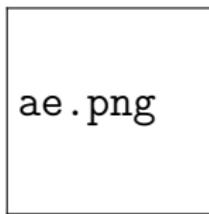
$$\hat{Y}(X) = F(X) = \left( f_1^{W_1, b_1} \circ \dots \circ f_L^{W_L, b_L} \right) (X)$$

# Deep Learning: Academic Curiosity? ... but it works so well!!

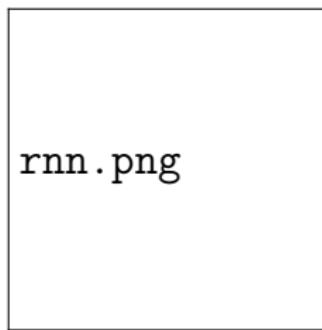
google-1.png

# Deep Architectures: TensorFlow

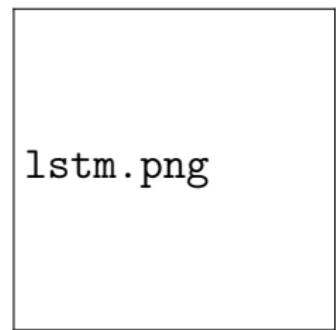
Many possible superpositions of univariate semi-affine functions



Auto-encoder



Recurrent



Long Short Term Memory

[http:](#)

[//www.asimovinstitute.org/neural-network-zoo/](http://www.asimovinstitute.org/neural-network-zoo/)

# Deep Learning Recovers Patterns

Matrix Algebra and Automatic Differentiation

# TPU Advantages for real-time data analysis

- ▶ Matrix analysis programmed into chips
- ▶ XLA: Accelerated Linear Algebra
- ▶ AD: Automatic Differentiation

# Area of Applications

- ▶ Image recognition: Healthcare, MNIST
- ▶ Sports
- ▶ Business
- ▶ ...

# Image recognition has improved

Machines are becoming better than humans

# Identifying Skin Cancer

- ▶ Dataset: 130,000 images of skin lesions/2,000 different diseases
- ▶ Test data: 370 high-quality, biopsy-confirmed images
- ▶ Better performance than 23 Stanford dermatologists
- ▶ 10,000 hours no match for deep learning and large datasets

# Google's AI: Heart disease from eye scan

Google's Verily scans of the back of a patient's eye able to accurately deduce individual's age, blood pressure, and whether or not they smoke, etc.

Predict their risk of suffering a major cardiac event—such as a heart attack—with roughly the same accuracy as current leading methods.

Quicker to analyze a patient's cardiovascular risk—doesn't require a blood test.

Training 300,000 patients. Eye scans plus general medical data.

Deep learning to mine for patterns

# Google's Verily



Two images of the fundus—interior rear of your eye.

The left is a regular image; the right shows how Google's algorithm picks out blood vessels (in green) to predict blood pressure.

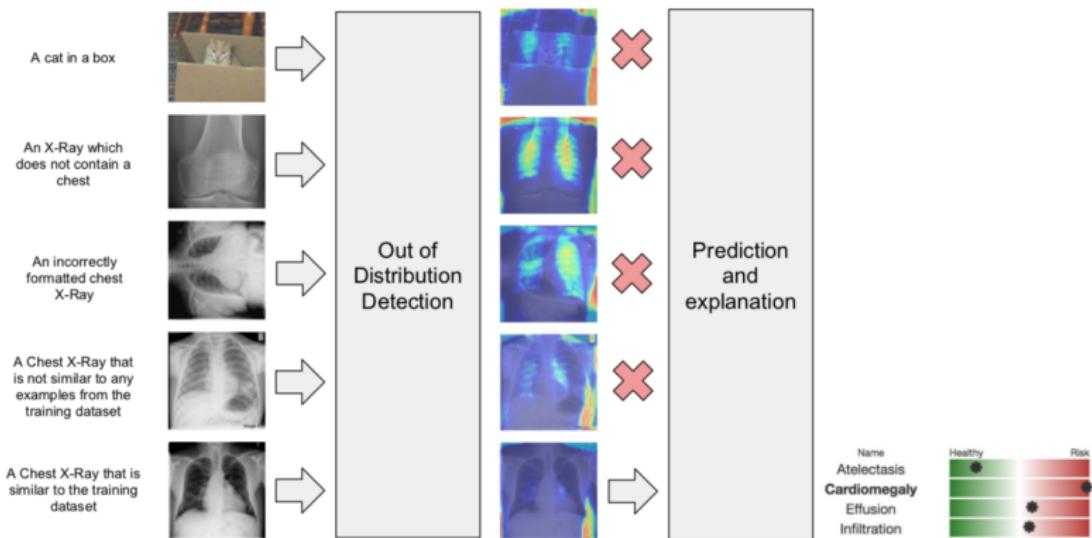
# Intelligent Scanning Using Deep Learning for MRI

As part of the MRI exam, the scan operator first acquires a set of low-resolution “localizer” images from which approximate location and orientation of desired landmarks can be identified. These anatomical references are then used to manually plan the exact locations, orientation, and required coverage for images that will be used for the high-resolution scans that are used for diagnosis.

To aid the scan operator we developed a deep-learning (DL) based framework for intelligent MRI slice placement (ISP) for several commonly used brain landmarks. TensorFlow library with the Keras interface is used to implement the DL based framework for ISP.

As compared to the classical approaches, a DL-based approach is less affected by factors that affect MRI image quality or appearance. And it can be easily extended across other anatomies.

# Chester: A Web Delivered Locally Computed Chest X-Ray Disease Prediction System



# Google Tulip

Google Assistant on phones and Google Home is now able to understand tulips, allowing translation between Tulipish and dozens of human languages.

Scientists have found that plants use their roots to send signals to neighboring plants, as a means to maintain their security and wellbeing.

The ability to speak with tulips comes with great environmental and societal benefits. Tulips now have a way to indicate to humans that they're in need of water, light or simply some more space. As their needs are expressed more clearly, they are able to live a happier and healthier life.

# Image Processing: MNIST

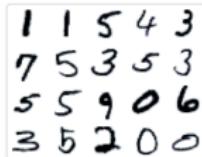
Hand-written digits. Multi-layer fully-connected neural network.  
Convolution neural network

**Figure:** Sample from MNIST datasets

# Classification Dataset Results

## MNIST

who is the best in MNIST ?



**MNIST** 50 results collected

Units: error %

Classify handwritten digits. Some additional results are available on the [original dataset page](#).

Result	Method	Venue	Details
0.21%	<a href="#">Regularization of Neural Networks using DropConnect</a>	ICML 2013	
0.23%	<a href="#">Multi-column Deep Neural Networks for Image Classification</a>	CVPR 2012	
0.23%	<a href="#">APAC: Augmented PAttern Classification with Neural Networks</a>	arXiv 2015	
0.24%	<a href="#">Batch-normalized Maxout Network in Network</a>	arXiv 2015	<a href="#">Details</a>
0.29%	<a href="#">Generalizing Pooling Functions in Convolutional Neural Networks: Mixed, Gated, and Tree</a>	AISTATS 2016	<a href="#">Details</a>
0.31%	<a href="#">Recurrent Convolutional Neural Network for Object Recognition</a>	CVPR 2015	
0.31%	<a href="#">On the Importance of Normalisation Layers in Deep Learning with Piecewise Linear Activation Units</a>	arXiv 2015	

# DL Model

The dataset contains 60k training observations and 10k samples out-of-sample performance (validation).

2-hidden layer MLP model with ReLU activation function

$$Y = \text{softmax}(W^3 \tilde{Y}_i^{(3)} + b^3)$$

$$\tilde{Y}_i^{(3)} = D^{(3)} \star Z^2$$

$$D_i^{(3)} \sim \text{Ber}(0.5)$$

$$Z^2 = \max(W^2 \tilde{Y}_i^{(2)} + b^1, 0)$$

$$\tilde{Y}_i^{(2)} = D^{(2)} \star Z^1$$

$$D_i^{(2)} \sim \text{Ber}(0.5)$$

$$Z^1 = \max(W^1 \tilde{Y}_i^{(1)} + b^1, 0)$$

$$\tilde{Y}_i^{(1)} = D^{(l)} \star X, \quad X \in R^{1024}$$

$$D_i^{(1)} \sim \text{Ber}(0.5)$$

Cross-entropy–negative log-likelihood–as loss function to be optimized.

# SGD and optimisation

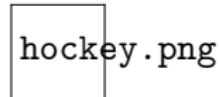
Apply mini-batch SGD algorithm with Nesterov acceleration (momentum) with the following parameters

- ▶ initial learning rate = 0.01
- ▶ learning decayed each time validation performance stalls (divided by 2)
- ▶ momentum of 0.9
- ▶ batch size of 10
- ▶ L2 weight decay / Gaussian prior on all parameters =  $1e-5$

# Sports!

# Learning Player Trajectories: NHL, NBA and EPL

Characteristics of group dynamics from their trajectories alone.



# Racehorse Big Data Unlocks the Formula for Human Superathletes

Jeff Seder had spent decades and millions of dollars collecting a huge database of physiological and biological data in an effort to discover which traits corresponded most closely with greatness. He pioneered portable ultrasound device that allowed him to examine horses on the inside and scanned tens of thousands of animals. It was only through this that he came to the conclusion that one of the most important data points in selecting a horse is the size of its heart – and American Pharoah had a huge one.

The wisdom of big data is increasingly being applied to human competition, although the field most certainly has room to grow. It will likely change not only how athletes perform, but how they look, as the tools and technologies of sports medicine become more and more sophisticated.

# Homecourt Can Help You Improve Your Jump Shot

HomeCourt is a basketball training app that uses deep learning to record, track, and chart shots for basketball players in real time. Using NVIDIA Tesla P100 GPUs on the Google Cloud, with the cuDNN-accelerated TensorFlow and Keras deep learning framework, the team trained their neural network on hours of basketball footage they filmed in local Bay Area high schools.

Not only can you review their workout videos with instant stat analysis but also workouts from other players and engage, and interact with the broader basketball community through basketball on a device they have with them all the time.

# Homecourt

- ▶ No sensors.

HomeCourt is ready to use with iPhone for real-time shot tracking and analysis.

- ▶ Interactive drills.

Your phone is transformed into your own virtual on-demand skills coach, with mobile AI-powered assessment tools.

- ▶ Shot Science.

HomeCourt's proprietary Shot Science technology provides meaningful insights for every shot you take.

- ▶ Analyze every shot.

With data tracked from >10M shots, >9M dribbles from >150 countries.

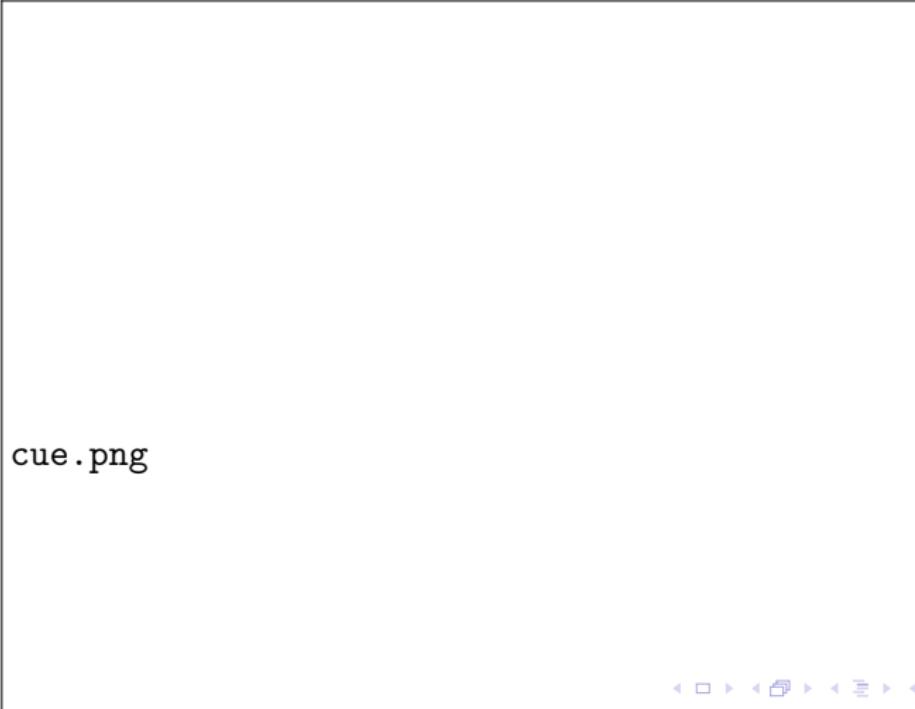
# Artificial Intelligence in Formula 1

## Formula One

- ▶ Strategy teams at the race track and at the team's HQ are constantly trying to predict the next best optimal move to improve their drivers' positions.
- ▶ Teams are limited to 60 data scientists  
AI (a.k.a machine learning/deep learning) provides better predictions of when best to stop, when to change tyres, overtake, ...
- ▶ Best strategies can vary quickly from moment to moment.

# Robot Shoot Hoops

This basketball-loving robot by Toyota played against pro ballers in a shooting competition. According to its creators, the bizarre-looking humanoid robot called 'Cue' learned how to shoot thanks to artificial intelligence.



# Google Data Center Cooling Costs Reduced by 40%

Monitoring real-time conditions and adjusting data center climate control based on past experience

Traffic and Driveless car.

# Traffic Prediction

Google maps real-time travel predictions ...

- ▶ Path search algorithms to calculate fastest route

# When Iowa's snow piles up, TensorFlow can keep roads safe

To improve road safety and efficiency, the Iowa Department of Transportation has teamed up with researchers at Iowa State University to use machine learning, including our TensorFlow framework, to provide insights into traffic behavior. Iowa State's technology helps analyze the visual data gathered from stationary cameras and cameras mounted on snow plows.

They also capture traffic information using radar detectors. Machine learning transforms that data into conclusions about road conditions, like identifying congestion and getting first responders to the scenes of accidents faster..

In California, snow may not be an issue, but traffic certainly is, and college students there used TensorFlow to identify pot holes and dangerous road cracks in Los Angeles.

# Stanford Cart

## Cart

- ▶ Hans Moravec and Stanford Cart
- ▶ KL10 processor, at about 2.5 MIPS, Moravec was eventually able to use multi-ocular vision to navigate slowly around obstacles in a controlled environment. The cart moved in one meter spurts punctuated by ten to fifteen minute pauses for image processing and route planning. In 1979, the cart successfully crossed a chair-filled room without human intervention in about five hours.

# Tesla deploys massive new Autopilot neural net in v9

Based on the new capabilities of Autopilot under version 9, the new computer vision neural net had to be significantly updated.

It can now track vehicles and other objects all around the car – meaning that it makes better use of the 8 cameras around the car and not just the front-facing ones.

Scaling computational power, training data, and industrial resources plays to Tesla's strengths and involves less uncertainty than potentially more powerful but less mature techniques.

# 1/8th-scale AWS DeepRacer machine-learning car

Amazon announced today that it will sell the AWS DeepRacer, a 1/8th-scale car with machine-learning technologies onboard.

Tinkerers can program the artificial-intelligence racers and, starting next year, will be able to compete in real-world races called the DeepRacer League.

The car's software uses reinforcement learning, where the car's software learns "on a trial-and-error basis within an interactive environment." You can use the Amazon SageMaker programming software to change the goals and rules for the DeepRacer's driving software then simulate it in some virtual race tracks within Amazon's software, before putting the car to use in real-world driving.

# Waymo: 5 billion hours of autonomous driving simulations

Alphabet-owned Waymo uses AI to run simulations in gaming-like environments. These simulations are being used to inform their autonomous vehicle efforts and the company has amassed an impressive 5 billion hours in edge case scenarios.

These edge cases are the really tricky 1% or 0.1% or even 0.001% of possible interactions that could occur when your vehicles move through complex environments.

While Waymo have some actual cars, it's clear their emphasis on safe simulations digitally, is the complete opposite to Tesla's strategy.

The big difference is that Tesla is deploying Autopilot into real customer vehicles and taking the data from events where the humans have taken over, therefore identifying an edge case that wasn't already accommodated in the system.

Business.

# Automated Rotterdam Port

## Port

Rotterdam Port is one of the most automated ports and one of the largest ports in the world.

Automated container carriers are completely computer controlled, carrying containers to cranes. Meanwhile, the cranes are human controlled and move the containers to the ship.

With the fully automated cranes, the terminal can be run by a team of no more than 10 to 15 people on a day-to-day basis.

# Automated Port: Port of Qingdao

## Qingdao

- ▶ Port of Qingdao is the first automated container terminal in Asia. The terminal is called a "ghost port" since it is all controlled by AI and no workers found in sight.
- ▶ Through laser scanning and positioning, the program is able to locate the four corners of each container. It accurately grabs them and puts them onto the driverless trucks. And it is capable to work in complete darkness. The smart autopilot trucks, driven by electricity, have their routes and tasks under digital control. They even know when it's time to go for a recharge.

# Rio Tinto Mining Automation to Boost Efficiency

## Rio Tinto

- ▶ 73 self-driving trucks that reportedly haul payloads at a cost 15 percent less than those operated by human drivers. In addition to the trucks, they also have robotic, rock-drilling rigs plugging away at the topography. In the near future. Rio Tinto is looking to upgrade the trains that haul the ore to port to not only drive themselves but also have the ability to load and unload automatically.
- ▶ 15% reduction in the cost of operating the automated trucks compared to those driven by humans, as hauling is among the largest costs to a mining operation.

# Applying deep learning to Related Pins

One of the most popular ways people find ideas on Pinterest is through Related Pins, an item-to-item recommendations system that uses collaborative filtering.

Previously, candidates were generated using board co-occurrence, signals from all the boards a Pin is saved to. Now, for the first time, Pinterest is applying deep learning to make Related Pins even more relevant. Ultimately, they developed a scalable system that evolves with their product and people's interests, the most relevant recommendations can surface through Related Pins.

Pin2Vec is built to embed all the Pins in a 128-dimension space. Pin tuples are used in supervised training to train the embedding matrix for each of the tens of millions of Pins of the vocabulary. TensorFlow is used as the trainer. At serving time, a set of nearest neighbors are found as Related Pins in the space for each of the Pins.

# Google DeepMind's Deep Q-learning playing Atari Breakout

Google DeepMind

- ▶ Google DeepMind created an AI using deep reinforcement learning that plays Atari games and improves itself to a superhuman level.
- ▶ Capable of playing many Atari games and uses a combination of deep artificial neural networks and reinforcement learning.
- ▶ This was the beginning for Google DeepMind.

OpenAI and Dota2 is the current state-of-the-art

# Watson and Jeopardy

## Jeopardy

- ▶ IBM super computer Watson beat two former champions of TV game jeopardy and took home one million dollars prize. Watson is a significant leap a machine's ability to understand context in human language.
- ▶ IBM believes the technology behind Watson can be applied to a variety of fields, most notably medicine.

# Google: $\alpha$ Go

## alphaGo Movie

Supervised and Reinforcement Learning, Value Function and Tree Search

### Convenient

- ▶ Fully observed
- ▶ Discrete action space
- ▶ Perfect simulator
- ▶ Relatively short game
- ▶ Trial-and error experience
- ▶ Large human datasets

### Inconvenient

- ▶ Actions executed awkwardly
- ▶ Incomplete information
- ▶ Imperfect simulator
- ▶ Longer tasks, hard to assess value
- ▶ Hard to practice millions of times
- ▶ Small human data sources

# AlphaZero's Groundbreaking Chess Strategies and the Promise of AI

How AlphaZero works – the theory

1. Learning rather than being programmed
2. General rather than specific
3. Grounded rather than logic-based
4. Active rather than passive

# Not HotDog

Finding Architecture Hard

Silicon Valley: Season 4 Episode 4:  
<https://youtu.be/ACmydtFDTGs>

# Topics: 41901

W1-W2: Probability and Bayes

W3-W4: Distributions and Expectations

W5: Modern Regression Methods (Lasso and Ridge)

W6-W7: Bayesian Hierarchical Models

W8: Brownian Motion

W9-W10: AI and Deep Learning

**HAVE FUN!!**