data science @ NYT

Chris Wiggins

Aug 8/9, 2016
1. overview of DS@NYT
Outline

1. overview of DS@NYT
2. prediction + supervised learning
Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
4. description + inference
Outline

1. overview of DS@NYT
2. prediction + supervised learning
3. prescription, causality, and RL
4. description + inference
5. (if interest) designing data products
0. Thank the organizers!

Figure 1: prepping slides until last minute
Lecture 1: overview of ds@NYT
data science @ The New York Times

chris.wiggins@columbia.edu
chris.wiggins@nytimes.com
@chrishwiggins

data science
data science: searches
data science: mindset & toolset

Drew Conway, 2010
Information Platforms and the Rise of the Data Scientist

Jeff Hammerbacher

Beautiful Data

The Stories Behind Elegant Data Solutions

Edited by Toby Segaran & Jeff Hammerbacher

modern history:
2009
At Facebook, we felt that traditional titles such as Business Analyst, Statistician, Engineer, and Research Scientist didn’t quite capture what we were after for our team. The workload for the role was diverse: on any given day, a team member could author a multistage processing pipeline in Python, design a hypothesis test, perform a regression analysis over data samples with R, design and implement an algorithm for some data-intensive product or service in Hadoop, or communicate the results of our analyses to other members of the organization in a clear and concise fashion. To capture the skill set required to perform this multitude of tasks, we created the role of “Data Scientist.”
biology: 1892 vs. 1995
biology: 1892 vs. 1995

biology changed for good.
biology: 1892 vs. 1995

new toolset, new mindset
genetics: 1837 vs. 2012

ML toolset; data science mindset
genetics: 1837 vs. 2012

Statistical Modeling: The Two Cultures
Leo Breiman
genetics: 1837 vs. 2012
ML toolset; data science mindset

arxiv.org/abs/1105.5821 ; github.com/rajanil/mkboost
data science: mindset & toolset
news: 20th century

church

state
The Best and Worst Places to Grow Up: How Your Area Compares

Children who grow up in some places go on to earn much more than they would if they grew up elsewhere.

08540

GENDER
All kids Boys Girls

INCOME PERCENTILE
25th 50th 75th

church

If a child in a poor family were to grow up in Suffolk County, Mass., instead of an average place, he or she would make $840, or 3 percent, less at age 26.
2014 The Year in Interactive Storytelling, Graphics and Multimedia

From a ship in the South China Sea to the cost of health care in the United States, the range of subjects here is broad, but the common thread is the form of storytelling — an integration of text, video, photography and graphics.
news: 20th century

church

state
news: 21st century

church

state

data
newspapering: 1851 vs. 1996

1851

The New York Times Introduces a Web Site
By Peter H. Lewis
Published: January 22, 1996

The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper's contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day's printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper's crossword puzzle.

1996
1,615,934 site-wide views over the last hour
1,257,958 average Sunday New York Times print circulation
554 stories written over the last 24 hours
206 countries with visitors in the past 25 minutes
243,192 words written in the last 24 hours
65 New York Times newspaper print sites globally
733 page views from India in the last 10 minutes
Blue Bottle Coffee

Trendy cafe chain offering upscale coffee drinks & pastries, plus beans & brewing equipment.

📍 160 Berry St, Brooklyn, NY 11249
⏰ Open today 7:00 am – 7:00 pm

Menu

📞 (510) 653-3394

Popular times: Tuesdays
"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great"
"...social activities generate large quantities of potentially valuable data...The data were not generated for the purpose of learning; however, the potential for learning is great" - J Chambers, Bell Labs, 1993, "GLS"
data science: the web
data science: the web

is your “online presence”
data science: the web

is a microscope
data science: the web

is an experimental tool
data science: the web

is an optimization tool
newspapering: 1851 vs. 1996 vs. 2008

The New York Times Introduces a Web Site

BY PETER H. LEWIS
Published: January 22, 1996

The New York Times begins publishing daily on the World Wide Web today, offering readers around the world immediate access to most of the daily newspaper’s contents.

The New York Times on the Web, as the electronic publication is known, contains most of the news and feature articles from the current day’s printed newspaper, classified advertising, reporting that does not appear in the newspaper, and interactive features including the newspaper’s crossword puzzle.
“a startup is a temporary organization in search of a repeatable and scalable business model” —Steve Blank
every publisher is now a startup
every publisher is now a startup
every publisher is now a startup

Advertisers adjusted spending accordingly. In the first quarter of 2016, 85 cents of every new dollar spent in online advertising will go to Google or Facebook, said Brian Nowak, a Morgan Stanley analyst.
3. Over time, disruptors improve their product, usually by adapting a new technology. The flashpoint comes when their products become “good enough” for most customers.

They are now poised to grow by taking market share from incumbents.

HALLMARKS OF DISRUPTIVE INNOVATORS

- Introduced by an “outsider”
- Less expensive than existing products
- Targeting underserved or new markets
- Initially inferior to existing products
- Advanced by an enabling technology
news: 21st century

curch

state

data
news: 21st century

church

state

data
learnings
learnings

- predictive modeling
- descriptive modeling
- prescriptive modeling
(actually ML, shhhh...) 

- (supervised learning) 
- (unsupervised learning) 
- (reinforcement learning)
h/t michael littman
Analytic Value Escalator

- Supervised Learning
- Reinforcement Learning
- Unsupervised Learning

h/t michael littman
learnings

- predictive modeling
- descriptive modeling
- prescriptive modeling

cf. modelingssocialdata.org
from “are you a bayesian or a frequentist”
—michael jordan
predictive modeling, e.g.,

cf. modelingsocialdata.org
predictive modeling, e.g.,

“the funnel”

cf. modelingsocialdata.org
interpretable predictive modeling

super cool stuff

Most Predictive Features

cf. modelingsocialdata.org
<table>
<thead>
<tr>
<th>TFNAME</th>
<th>DB-MOTIF</th>
<th>MOTIF</th>
<th>DBNAME</th>
<th>d(p,q)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBF1</td>
<td>CACCTG</td>
<td>CACCTG</td>
<td>YPD</td>
<td>0.029285</td>
</tr>
<tr>
<td>CCG eveted repeat</td>
<td>CCGN*CCG</td>
<td>CCGN*CCG</td>
<td>YPD</td>
<td>0.032921</td>
</tr>
<tr>
<td>MSN2</td>
<td>AGCCGG</td>
<td>AGCCGG</td>
<td>TRANSFAC</td>
<td>0.085626</td>
</tr>
<tr>
<td>HSF1</td>
<td>TTCNNNGAA</td>
<td>TTCNNNGAA</td>
<td>SCPD</td>
<td>0.102410</td>
</tr>
<tr>
<td>XBP1</td>
<td>TCGAAC</td>
<td>TCGAAC</td>
<td>TRANSFAC</td>
<td>0.140561</td>
</tr>
<tr>
<td>STE12</td>
<td>TGAAC</td>
<td>TGAAC</td>
<td>TRANSFAC</td>
<td>0.296730</td>
</tr>
<tr>
<td>GCN4</td>
<td>TGAAC</td>
<td>TGAAC</td>
<td>SCPD</td>
<td>0.292221</td>
</tr>
<tr>
<td>TBP</td>
<td>TTTATA</td>
<td>TTTATA</td>
<td>TRANSFAC</td>
<td>0.376801</td>
</tr>
<tr>
<td>HAP1</td>
<td>CGGNNNTWNCGG</td>
<td>CGGNNNTWNCGG</td>
<td>YPD</td>
<td>0.429004</td>
</tr>
<tr>
<td>RAP1</td>
<td>RMACCCA</td>
<td>RMACCCA</td>
<td>SCPD</td>
<td>0.529059</td>
</tr>
<tr>
<td>mPAC</td>
<td>AAAAAATT</td>
<td>AAAAAATT</td>
<td>AlignACE</td>
<td>0.552493</td>
</tr>
<tr>
<td>mRRPE</td>
<td>AAAAAATT</td>
<td>AAAAAATT</td>
<td>AlignACE</td>
<td>0.630740</td>
</tr>
<tr>
<td>PHO4</td>
<td>CACGT</td>
<td>CACGT</td>
<td>TRANSFAC</td>
<td>0.672961</td>
</tr>
<tr>
<td>YAP1</td>
<td>TATG</td>
<td>TATG</td>
<td>TRANSFAC</td>
<td>0.777816</td>
</tr>
<tr>
<td>MIG1</td>
<td>CCCACAAA</td>
<td>CCCACAAA</td>
<td>YPD</td>
<td>0.799412</td>
</tr>
<tr>
<td>MET31.32</td>
<td>AACTGTGG</td>
<td>AACTGTGG</td>
<td>YPD</td>
<td>0.84893</td>
</tr>
<tr>
<td>HAP2,3,4</td>
<td>CCAT</td>
<td>CCAT</td>
<td>TRANSFAC</td>
<td>1.076837</td>
</tr>
</tbody>
</table>

arxiv.org/abs/q-bio/0701021
optimization & learning, e.g.,

optimization & prediction, e.g.,

“newsvendor problem,” literally (+prediction+experiment)
recommendation as inference

<table>
<thead>
<tr>
<th>MOST EMAILED</th>
<th>MOST VIEWED</th>
<th>RECOMMENDED FOR YOU</th>
</tr>
</thead>
</table>
| 1. THE OUTLAW OCEAN  
A Renegade Trawler, Hunted for 10,000 Miles by Vigilantes | ![Image](image1) | ![Image](image2) |
| 2. Campus Suicide and the Pressure of Perfection | ![Image](image3) | ![Image](image4) |
| 3. As Tech Booms, Workers Turn to Coding for Career Change | ![Image](image5) | ![Image](image6) |
| 4. Prison Worker Who Aided Escape Tells of Sex, Saw Blades and Deception | ![Image](image7) | ![Image](image8) |
| 5. Under Oath, Donald Trump Shows His Raw Side | ![Image](image9) | ![Image](image10) |
| 6. American Hunter Killed Cod, Beloved Lion That Was Lured Out of Its Sanctuary | ![Image](image11) | ![Image](image12) |
| 7. A Creature on the Loose Puts Milwaukee Residents on Edge | ![Image](image13) | ![Image](image14) |
| 8. N.F.L. Upholds Tom Brady’s Ban; Cellphone’s Fate Helped Make the Call | ![Image](image15) | ![Image](image16) |
| 9. Escalator Death in China Sets Off Furor Online | ![Image](image17) | ![Image](image18) |
| 10. DAVID BROOKS  
The Structure of Gratitude | ![Image](image19) | ![Image](image20) |
recommendation as inference
descriptive modeling, e.g.,

“segments”

cf. modelingsocialdata.org
descriptive modeling, e.g.,

“segments”

cf. modelingsocialdata.org
descriptive modeling, e.g.,

\[
\text{argmax}_z \ p(z|x) = 14
\]

“segments”

cf. modelingsocialdata.org
descriptive modeling, e.g.,

“segments”

“baby boomer”

cf. modelingsocialdata.org
- descriptive data product

A Quick Refinery Demo

Extracting NYT articles from keyword “obama” in 2013.

What themes / topics defined the Obama administration during 2013?

cf. daeilkim.com
descriptive modeling, e.g.,

cf. daeilkim.com ; import bnpy
modeling your audience

bit.ly/Hughes-Kim-Sudderth-AISTATS15
Objective function. Mean field methods optimize an evidence lower bound \( \log p(x|\gamma, \alpha, \tau) \geq \mathcal{L}(\cdot) \), where

\[
\mathcal{L}(\cdot) \triangleq \mathcal{L}_{data}(\cdot) + H_z(\cdot) + \mathcal{L}_{HDP}(\cdot) + \mathcal{L}_u(\cdot). \quad (4)
\]

The final term \( \mathcal{L}_u(\cdot) \), which depends only on \( q(u) \), is discussed in the next section. The first three terms account for data generation, the assignment entropy, and the document-topic allocations. These are defined below, with expectations taken with respect to Eq. (3):

\[
\mathcal{L}_{data}(\cdot) \triangleq \mathbb{E}_q[\log p(x|z, \phi) + \log \frac{p(\phi|\tau)}{q(\phi|\tau)}], \quad (5)
\]

\[
H_z(\cdot) \triangleq - \sum_{k=1}^{K} \sum_{d=1}^{D} \sum_{n=1}^{N_d} \hat{r}_{dnk} \log \hat{r}_{dnk},
\]

\[
\mathcal{L}_{HDP}(\cdot) \triangleq \mathbb{E}_q \left[ \log \frac{p(z|\pi)p(\pi|\alpha, \tau)}{q(\pi|\tilde{\theta})} \right].
\]

The forms of \( \mathcal{L}_{data} \) and \( H_z \) are unchanged from the simpler case of mean-field for DP mixtures. Closed-form expressions are in the Supplement.
modeling your audience
also allows insight+targeting as inference
prescriptive modeling
**Prescriptive Modeling**

<table>
<thead>
<tr>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>Specify $x$; learn $z(x)$ or $p(z</td>
</tr>
<tr>
<td>Predictive</td>
<td>Specify $x$ and $y$; learn to predict $y$ from $x$</td>
</tr>
<tr>
<td>Prescriptive</td>
<td>Specify $x$, $y$, and $a$; learn to prescribe $a$ given $x$ to maximize $y$</td>
</tr>
</tbody>
</table>
prescriptive modeling

\[ V = E_+(y) = \sum_{yax} yP_+(y, a, x) \]
“off policy value estimation”
(cf. “causal effect estimation”)

\[
\hat{V} = \frac{1}{N} \sum_{i=1}^{i=N} y_i \frac{1(a_i = h(x_i))}{\hat{B}(a_i|x_i)}
\]

cf. Langford `08-`16;
Horvitz & Thompson `52;
Holland `86
“off policy value estimation” (cf. “causal effect estimation”)

\[
\hat{V} = \frac{1}{N} \sum_{i=1}^{i=N} y_i \frac{1(a_i = h(x_i))}{\hat{B}(a_i|x_i)}
\]

Vapnik’s razor
“When solving a (learning) problem of interest, do not solve a more complex problem as an intermediate step.”
prescriptive modeling

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer\textsuperscript{a,1}, Jamie E. Guillory\textsuperscript{b,2}, and Jeffrey T. Hancock\textsuperscript{b,3}

Author Affiliations

\textsuperscript{a}Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and
\textsuperscript{b}Departments of Communication and
\textsuperscript{c}Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published
A correction has been published

cf. modelingsocialdata.org
prescriptive modeling

Experimental evidence of massive-scale emotional contagion through social networks

Adam D. I. Kramer\textsuperscript{a,1}, Jamie E. Guillory\textsuperscript{b,2}, and Jeffrey T. Hancock\textsuperscript{b,3}

Author Affiliations

\textsuperscript{a}Core Data Science Team, Facebook, Inc., Menlo Park, CA 94025; and Departments of \textsuperscript{b}Communication and \textsuperscript{c}Information Science, Cornell University, Ithaca, NY 14853

Edited by Susan T. Fiske, Princeton University, Princeton, NJ, and approved March 25, 2014 (received for review October 23, 2013)

A correction has been published
A correction has been published

aka “A/B testing”; RCT

cf. modelingsocialdata.org
Some of the most recognizable personalization in our service is the collection of “genre” rows. ...Members connect with these rows so well that we measure an increase in member retention by placing the most tailored rows higher on the page instead of lower.

cf. modelingsocialdata.org
real-time A/B -> “bandits”

GOOG blog:

cf. modelingsocialdata.org
prescriptive modeling, e.g,
prescriptive modeling, e.g,
prescriptive modeling, e.g,
prescriptive modeling, e.g.,

leverage methods which are predictive yet performant
NB: data-informed, not data-driven
predicting views/cascades: doable?

KDD 09: how many people are online?
predicting views/cascades: features?

<table>
<thead>
<tr>
<th>Content Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>score_food/nature/...</td>
</tr>
<tr>
<td>is_en</td>
</tr>
<tr>
<td>has_caption</td>
</tr>
<tr>
<td>is_re/eng</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Root (Original Poster) Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
</tr>
<tr>
<td>orig_is_page</td>
</tr>
<tr>
<td>outdeg(v)</td>
</tr>
<tr>
<td>agei &amp;</td>
</tr>
<tr>
<td>genderi</td>
</tr>
<tr>
<td>fbb_age</td>
</tr>
<tr>
<td>activity</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Resharer Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>src</td>
</tr>
<tr>
<td>pages</td>
</tr>
<tr>
<td>friends</td>
</tr>
<tr>
<td>fans</td>
</tr>
<tr>
<td>followers</td>
</tr>
<tr>
<td>birthdate</td>
</tr>
<tr>
<td>activities</td>
</tr>
<tr>
<td>age</td>
</tr>
<tr>
<td>female</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Structural Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>outdeg(v_i)</td>
</tr>
<tr>
<td>outdeg(v_i)</td>
</tr>
<tr>
<td>outdeg(v_i)</td>
</tr>
<tr>
<td>orig_con</td>
</tr>
<tr>
<td>border_nodes</td>
</tr>
<tr>
<td>border_edges</td>
</tr>
<tr>
<td>subgraph</td>
</tr>
<tr>
<td>depth</td>
</tr>
<tr>
<td>depth</td>
</tr>
<tr>
<td>did</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Temporal Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>time</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>time</td>
</tr>
<tr>
<td>time</td>
</tr>
</tbody>
</table>
predicting views/cascades: features?

Figure 10: In predicting the largest cascade in clusters of 10 or more cascades of identical photos, we perform significantly above the baseline of 0.1.
predicting views/cascades: doable?

Exploring limits to prediction in complex social systems

Travis Martin
University of Michigan
Dept. of Computer Science
Ann Arbor, MI
travisbm@umich.edu

Jake M. Hofman
Microsoft Research
641 6th Ave, Floor 7
New York, NY
jmh@microsoft.com

Amit Sharma
Microsoft Research
amshar@microsoft.com

Ashton Anderson
Microsoft Research
ashton@microsoft.com

Duncan J. Watts
Microsoft Research
duncan@microsoft.com

WWW 16: TWIT RT’s
predicting views/cascades: doable?

Figure 4: Prediction results for models using different subsets of features. $R^2$ increases as we add more features, but only up to a limit. Even a model with all features explains less than half of the variance in cascade sizes.
descriptive:

predictive:

prescriptive:

Explore
↓
Learning
↓
Test
↓
Optimizing
↓
Reporting
things:
what does DS team deliver?

- build data product
- build APIs
- impact roadmaps
data science @ The New York Times

chris.wiggins@columbia.edu
chris.wiggins@nytimes.com
@chrishwiggins

Lecture 2: predictive modeling @ NYT
<table>
<thead>
<tr>
<th>Type</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>Descriptive</td>
<td>specify $x$; learn $z(x)$ or $p(z</td>
</tr>
<tr>
<td>Predictive</td>
<td>specify $x$ and $y$; learn to predict $y$ from $x$</td>
</tr>
<tr>
<td>Prescriptive</td>
<td>specify $x$, $y$, and $a$; learn to prescribe $a$ given $x$ to maximize $y$</td>
</tr>
</tbody>
</table>

Figure 2: desc/pred/pres

- caveat: difference between observation and experiment. why?
blossom example

prescriptive modeling, e.g.,

blossombot: BOF 3:48 PM ⭐
Blossom Alert!
This piece is predicted to go viral if posted on Facebook next:
Here is its staLink


leverage methods which are predictive yet performant

Figure 3: Reminder: Blossom
margin-based surrogate loss functions

\[ L = \sum_{i=1}^{N} \varphi(y_i f(x_i; \beta)) + \lambda \| \beta \| \]

from “are you a bayesian or a frequentist”
- Michael Jordan

Figure 4: Reminder: Surrogate Loss Functions
tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in \mathbb{R}$
tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log \frac{p(y = 1|x)}{p(y = -1|x)} \in \mathbb{R}$
- $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = \frac{1}{1 + \exp(-yf)}$
tangent: logistic function as surrogate loss function

- Define $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- $-\log_2 p(\{y\}^N_1) = \sum_i \log_2 \left(1 + e^{-y_if(x_i)}\right) \equiv \sum_i \ell(y_if(x_i))$
tangent: logistic function as surrogate loss function

- define $f(x) \equiv \log p(y = 1|x)/p(y = -1|x) \in R$
- $p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$
- $-\log_2 p(\{y\}_1^N) = \sum_i \log_2 (1 + e^{-y_if(x_i)}) \equiv \sum_i \ell(y_if(x_i))$
- $\ell'' > 0, \ell(\mu) > 1[\mu < 0] \quad \forall \mu \in R$. 
define $f(x) \equiv \log \frac{p(y = 1|x)}{p(y = -1|x)} \in \mathbb{R}$

$p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = 1/(1 + \exp(-yf))$

$-\log_2 p(\{y\}_{1}^{N}) = \sum_{i} \log_2 \left(1 + e^{-y_i f(x_i)}\right) \equiv \sum_{i} \ell(y_i f(x_i))$

$\ell'' > 0, \ell(\mu) > 1[\mu < 0] \ \forall \mu \in \mathbb{R}$.

$\therefore$ maximizing log-likelihood is minimizing a surrogate convex loss function for classification (though not strongly convex, cf. Yoram’s talk)
define \( f(x) \equiv \log \frac{p(y = 1|x)}{p(y = -1|x)} \in \mathbb{R} \)

\[ p(y = 1|x) + p(y = -1|x) = 1 \rightarrow p(y|x) = \frac{1}{1 + \exp(-yf(x))} \]

\[-\log_2 p(\{y\}_1^N) = \sum_i \log_2 \left(1 + e^{-y_i f(x_i)}\right) \equiv \sum_i \ell(y_i f(x_i)) \]

\[ \ell'' > 0, \ell(\mu) > 1[\mu < 0] \quad \forall \mu \in \mathbb{R}. \]

\[ \therefore \text{maximizing log-likelihood is minimizing a surrogate convex loss function for classification (though not strongly convex, cf. Yoram’s talk)} \]

but \( \sum_i \log_2 \left(1 + e^{-y_i w^T h(x_i)}\right) \) not as easy as \( \sum_i e^{-y_i w^T h(x_i)} \)
$L$ exponential surrogate loss function, summed over examples:

\[ L[F] = \sum_i \exp (-y_i F(x_i)) \]
$L$ exponential surrogate loss function, summed over examples:

$$L[F] = \sum_i \exp (-y_i F(x_i))$$

$$= \sum_i \exp (-y_i \sum_{t'} w_{t'} h_{t'}(x_i)) \equiv L_t(w_t)$$
$L$ exponential surrogate loss function, summed over examples:

- $L[F] = \sum_i \exp(-y_i F(x_i))$
- $= \sum_i \exp(-y_i \sum_{t'} w_{t'} h_{t'}(x_i)) \equiv L_t(w_t)$
- Draw $h_t \in \mathcal{H}$ large space of rules s.t. $h(x) \in \{-1, +1\}$
$L$ exponential surrogate loss function, summed over examples:

$\triangleright \quad L[F] = \sum_i \exp(-y_i F(x_i))$

$\triangleright \quad = \sum_i \exp(-y_i \sum_{t'}^t w_t' h_t'(x_i)) \equiv L_t(w_t)$

$\triangleright \quad$ Draw $h_t \in \mathcal{H}$ large space of rules s.t. $h(x) \in \{-1, +1\}$

$\triangleright \quad$ label $y \in \{-1, +1\}$
$L$ exponential surrogate loss function, summed over examples:

$L_{t+1}(w_t; w) \equiv \sum_i d^t_i \exp (-y_i w h_{t+1}(x_i))$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, $L_1, L_\infty^1, \ldots$
$L$ exponential surrogate loss function, summed over examples:

\[
\begin{align*}
L_{t+1}(w_t; w) & \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i)) \\
& = \sum_{y=h'} d_i^t e^{-w} + \sum_{y\neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-
\end{align*}
\]

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, $L_1, L_\infty^1, \ldots$
$L$ exponential surrogate loss function, summed over examples:

$- L_{t+1}(w_t; w) \equiv \sum_i d_i^t \exp (-y_i w h_{t+1}(x_i))$

$= \sum_{y=h} d_i^t e^{-w} + \sum_{y\neq h} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$

$\therefore w_{t+1} = \arg\min_w L_{t+1}(w) = (1/2) \log D_+/D_-$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, $L_1, L_\infty^1, \ldots$
L exponential surrogate loss function, summed over examples:

\[
L_{t+1}(w_t; w) \equiv \sum_i d_i^t \exp(-y_i w h_{t+1}(x_i))
\]
\[
= \sum_{y=h'} d_i^t e^{-w} + \sum_{y \neq h'} d_i^t e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-
\]
\[
\therefore w_{t+1} = \arg\min_w L_{t+1}(w) = (1/2) \log D_+/D_-
\]
\[
L_{t+1}(w_{t+1}) = 2\sqrt{D_+D_-} = 2\sqrt{\nu_+(1-\nu_+)}/D,
\]
where
\[
0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1
\]

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, \( L_1, L_\infty \). . .
L exponential surrogate loss function, summed over examples:

- $L_{t+1}(w_t; w) \equiv \sum_i d_t^i \exp (-y_i w h_{t+1}(x_i))$
- $= \sum_{y=h'} d_t^i e^{-w} + \sum_{y \neq h'} d_t^i e^{+w} \equiv e^{-w} D_+ + e^{+w} D_-$
- $\therefore w_{t+1} = \arg\min_w L_{t+1}(w) = (1/2) \log D_+/D_-$
- $L_{t+1}(w_{t+1}) = 2\sqrt{D_+D_-} = 2\sqrt{\nu_+(1-\nu_+)} / D$, where $0 \leq \nu_+ \equiv D_+/D = D_+/L_t \leq 1$
- update example weights $d_{t+1}^i = d_t^i e^{\mp w}$

Punchlines: sparse, predictive, interpretable, fast (to execute), and easy to extend, e.g., trees, flexible hypotheses spaces, $L_1, L_\infty^1, \ldots$

---

$^1$Duchi + Singer “Boosting with structural sparsity” ICML ’09
predicting people

- “customer journey” prediction
predicting people

- “customer journey” prediction
  - fun covariates
predicting people

- “customer journey” prediction
  - fun covariates
  - observational complication v structural models
Figure 5: both in science and in real world, feature analysis guides future experiments
Figure 6: from Lecture 1
example in CAR (computer assisted reporting)

Figure 7: Tabuchi article
example in CAR (computer assisted reporting)

▶ cf. Friedman’s “Statistical models and Shoe Leather”\(^2\)

example in CAR (computer assisted reporting)

- cf. Friedman’s “Statistical models and Shoe Leather”\(^2\)
- Takata airbag fatalities

---

example in CAR (computer assisted reporting)

- cf. Friedman’s “Statistical models and Shoe Leather”\textsuperscript{2}
- Takata airbag fatalities
- 2219 labeled\textsuperscript{3} examples from 33,204 comments


\textsuperscript{3}By Hiroko Tabuchi, a Pulitzer winner
example in CAR (computer assisted reporting)

- cf. Friedman’s “Statistical models and Shoe Leather”\(^2\)
- Takata airbag fatalities
- 2219 labeled\(^3\) examples from 33,204 comments
- cf. Box’s “Science and Statistics”\(^4\)

---


\(^3\)By Hiroko Tabuchi, a Pulitzer winner

computer assisted reporting

- Impact

Figure 8: impact
Lecture 3: prescriptive modeling @ NYT
the natural abstraction

- operators\(^5\) make decisions

---

\(^5\)In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines
the natural abstraction

- operators\(^5\) make decisions
- faster horses v. cars

\(^5\)In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines
the natural abstraction

- operators\(^5\) make decisions
- faster horses v. cars
- general insights v. optimal policies

\(^5\)In the sense of business deciders; that said, doctors, including those who operate, also have to make decisions, cf., personalized medicines
maximizing outcome

- the problem: maximizing an outcome over policies...
maximizing outcome

- the problem: maximizing an outcome over policies...
- ...while inferring causality from observation
maximizing outcome

- the problem: maximizing an outcome over policies...
- ...while inferring causality from observation
- different from predicting outcome in absence of action/policy
examples

- observation is not experiment
examples

- observation is not experiment
  - e.g., (Med.) smoking hurts vs unhealthy people smoke
examples

- observation is not experiment
  - e.g., (Med.) smoking hurts vs unhealthy people smoke
  - e.g., (Med.) affluent get prescribed different meds/treatment
examples

- observation is not experiment
  - e.g., (Med.) smoking hurts vs unhealthy people smoke
  - e.g., (Med.) affluent get prescribed different meds/treatment
  - e.g., (life) veterans earn less vs the rich serve less

---

examples

- observation is not experiment
  - e.g., (Med.) smoking hurts vs unhealthy people smoke
  - e.g., (Med.) affluent get prescribed different meds/treatment
  - e.g., (life) veterans earn less vs the rich serve less\(^6\)
  - e.g., (life) admitted to school vs learn at school?

---

reinforcement/machine learning/graphical models

- key idea: model joint $p(y, a, x)$
key idea: model joint $p(y, a, x)$
explore/exploit: family of joints $p_\alpha(y, a, x)$
key idea: model joint $p(y, a, x)$
explore/exploit: family of joints $p_\alpha(y, a, x)$
“causality”: $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$ “a causes y”
key idea: model joint $p(y, a, x)$

explore/exploit: family of joints $p_\alpha(y, a, x)$

“causality”: $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$ “a causes y”

nomenclature: ‘response’, ‘policy’/‘bias’, ‘prior’ above
in general

Figure 9: policy/bias, response, and prior define the distribution also describes both the ‘exploration’ and ‘exploitation’ distributions
randomized controlled trial

Figure 10: RCT: ‘bias’ removed, random ‘policy’ (response and prior unaffected)

also Pearl’s ‘do’ distribution: a distribution with “no arrows” pointing to the action variable.
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
  - role of “IPW”
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
  - role of “IPW”
- reduction
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
  - role of “IPW”
- reduction
- normalization
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
  - role of “IPW”
- reduction
- normalization
- hyper-parameter searching
POISE: calculation, estimation, optimization

- POISE: “policy optimization via importance sample estimation”
- Monte Carlo importance sampling estimation
  - aka “off policy estimation”
  - role of “IPW”
- reduction
- normalization
- hyper-parameter searching
- unexpected connection: personalized medicine
POISE setup and Goal

> "a causes y" ⇐⇒ ∃ family \( p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x) \)
POISE setup and Goal

- “a causes y” $\iff \exists$ family $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$
- define off-policy/exploration distribution $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
POISE setup and Goal

- “a causes y” $\iff \exists$ family $p_{\alpha}(y, a, x) = p(y|a, x)p_{\alpha}(a|x)p(x)$
- define off-policy/exploration distribution
  $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$
- define exploitation distribution
  $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$
POISE setup and Goal

“a causes y” ⇐⇒ ∃ family \( p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x) \)

define off-policy/exploration distribution
\[ p_- (y, a, x) = p(y|a, x)p_-(a|x)p(x) \]

define exploitation distribution
\[ p_+ (y, a, x) = p(y|a, x)p_+(a|x)p(x) \]

Goal: Maximize \( E_+ (Y) \) over \( p_+(a|x) \) using data drawn from \( p_-(y, a, x) \).
POISE setup and Goal

- “a causes y” \iff \exists \text{ family } p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)
- define off-policy/exploration distribution
  \[ p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x) \]
- define exploitation distribution
  \[ p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x) \]
- Goal: Maximize \[ E_+(Y) \] over \[ p_+(a|x) \] using data drawn from \[ p_-(y, a, x) \].
POISE setup and Goal

- “a causes y” $\iff \exists$ family $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$

- define off-policy/exploration distribution
  $p_-(y, a, x) = p(y|a, x)p_-(a|x)p(x)$

- define exploitation distribution
  $p_+(y, a, x) = p(y|a, x)p_+(a|x)p(x)$

- Goal: Maximize $E_+(Y)$ over $p_+(a|x)$ using data drawn from $p_-(y, a, x)$.

notation: $\{x, a, y\} \in \{X, A, Y\}$ i.e., $E_\alpha(Y)$ is not a function of $y$
POISE math: $\text{IS} + \text{Monte Carlo estimation} = \text{ISE}$

i.e., “importance sampling estimation”

$E_+(Y) \equiv \sum_{yax} yp_+(y, a, x)$
POISE math: IS + Monte Carlo estimation = ISE

i.e., “importance sampling estimation”

\[ E_+(Y) \equiv \sum_{yax} yp_+(y, a, x) \]
\[ E_-(Y) = \sum_{yax} yp_-(y, a, x) (p_+(y, a, x)/p_-(y, a, x)) \]
i.e, “importance sampling estimation”

\[ E_+(Y) \equiv \sum_{yax} yp_{+}(y, a, x) \]
\[ E_+(Y) = \sum_{yax} yp_{-}(y, a, x) \left( p_{+}(y, a, x)/p_{-}(y, a, x) \right) \]
\[ E_+(Y) = \sum_{yax} yp_{-}(y, a, x) \left( p_{+}(a|x)/p_{-}(a|x) \right) \]
POISE math: IS+Monte Carlo estimation = ISE

i.e., “importance sampling estimation”

\[ E_+(Y) \equiv \sum_{yax} yp_+(y, a, x) \]
\[ E_+(Y) = \sum_{yax} yp_-(y, a, x) (p_+(y, a, x)/p_-(y, a, x)) \]
\[ E_+(Y) = \sum_{yax} yp_-(y, a, x) (p_+(a|x)/p_-(a|x)) \]
\[ E_+(Y) \approx N^{-1} \sum_i y_i (p_+(a_i|x_i)/p_-(a_i|x_i)) \]
POISE math: IS + Monte Carlo estimation = ISE

i.e., “importance sampling estimation”

- \[ E_+(Y) \equiv \sum_{yax} yp_+(y, a, x) \]
- \[ E_+(Y) = \sum_{yax} yp_-(y, a, x)(p_+(y, a, x)/p_-(y, a, x)) \]
- \[ E_+(Y) = \sum_{yax} yp_-(y, a, x)(p_+(a|x)/p_-(a|x)) \]
- \[ E_+(Y) \approx N^{-1} \sum_i y_i(p_+(a_i|x_i)/p_-(a_i|x_i)) \]
POISE math: IS+Monte Carlo estimation=ISE

i.e, “importance sampling estimation”

\[ E_+(Y) \equiv \sum_{y,a,x} yp_+(y, a, x) \]
\[ E_+(Y) = \sum_{y,a,x} yp_-(y, a, x) \left( \frac{p_+(y, a, x)}{p_-(y, a, x)} \right) \]
\[ E_+(Y) = \sum_{y,a,x} yp_-(y, a, x) \left( \frac{p_+(a|x)}{p_-(a|x)} \right) \]
\[ E_+(Y) \approx N^{-1} \sum_i y_i \left( \frac{p_+(a_i|x_i)}{p_-(a_i|x_i)} \right) \]

let’s spend some time getting to know this last equation, the importance sampling estimate of outcome in a “causal model” (“a causes y”) among \( \{y, a, x\} \)
Observation (cf. Bottou\textsuperscript{7} )

\[ \text{factorizing } P_\pm(x) \colon \frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{+ \text{but } \text{not}-}(x)}{P_{- \text{but } \text{not}+}(x)} \]
Observation (cf. Bottou\textsuperscript{7})

- factorizing $P_{\pm}(x)$: $\frac{P_{+}(x)}{P_{-}(x)} = \prod_{\text{factors}} \frac{P_{+\text{but not}-}(x)}{P_{-\text{but not}+}(x)}$
- origin: importance sampling $E_q(f) = E_p(fq/p)$ (as in variational methods)
Observation (cf. Bottou\textsuperscript{7})

- factorizing $P_{\pm}(x)$: $\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{\text{but not -}}(x)}{P_{\text{but not +}}(x)}$
- origin: importance sampling $E_q(f) = E_p(fq/p)$ (as in variational methods)
- the “causal” model $p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)$ helps here
Observation (cf. Bottou\(^7\))

- factorizing \(P_\pm(x)\): \(\frac{P_+(x)}{P_-(x)} = \prod_{\text{factors}} \frac{P_{\text{but not}}-+(x)}{P_{\text{but not}}-+(x)}\)
- origin: importance sampling \(E_q(f) = E_p(fq/p)\) (as in variational methods)
- the “causal” model \(p_\alpha(y, a, x) = p(y|a, x)p_\alpha(a|x)p(x)\) helps here
- factors left over are numerator \(p_+(a|x), \text{to optimize}\) and denominator \(p_-(a|x), \text{to infer if not a RCT}\)
Observation (cf. Bottou\textsuperscript{7})

- factorizing $P_{\pm}(x)$: \[ \frac{P_{+}(x)}{P_{-}(x)} = \prod_{\text{factors}} \frac{P_{\text{but not-}}(x)}{P_{\text{but not+}}(x)} \]
- origin: importance sampling $E_{q}(f) = E_{p}(f_{q}/p)$ (as in variational methods)
- the “causal” model $p_{\alpha}(y, a, x) = p(y|a, x)p_{\alpha}(a|x)p(x)$ helps here
- factors left over are numerator ($p_{+}(a|x)$, to optimize) and denominator ($p_{-}(a|x)$, to infer if not a RCT)
- unobserved confounders will confound us (later)

\textsuperscript{7}Counterfactual Reasoning and Learning Systems, arXiv:1209.2355
Reduction (cf. Langford\textsuperscript{8,9,10} (’05, ’08, ’09 ))

- consider numerator for deterministic policy:
  \[ p_+(a|x) = 1[a = h(x)] \]
Reduction (cf. Langford\textsuperscript{8,9,10} (’05, ’08, ’09))

- consider numerator for deterministic policy:
  \[ p_+(a|x) = 1[a = h(x)] \]
  \[ E_+(Y) \propto \sum_i (y_i/p_-(a|x))1[a = h(x)] \equiv \sum_i w_i1[a = h(x)] \]
Reduction (cf. Langford$^8,^9,^{10}$ ('05, '08, '09 ))

- consider numerator for deterministic policy:
  \[ p_+(a|x) = 1[a = h(x)] \]

- \[ E_+(Y) \propto \sum_i (y_i/p_-(a|x))1[a = h(x)] = \sum_i w_i 1[a = h(x)] \]

- Note: \[ 1[c = d] = 1 - 1[c \neq d] \]
Reduction (cf. Langford\textsuperscript{8,9,10} (’05, ’08, ’09 ))

- consider numerator for deterministic policy:
  \[ p_+(a|x) = 1[a = h(x)] \]
- \[ E_+(Y) \propto \sum_i (y_i / p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)] \]
- Note: \[ 1[c = d] = 1 - 1[c \neq d] \]
- \[ \therefore E_+(Y) \propto \text{constant} - \sum_i w_i 1[a \neq h(x)] \]
Reduction (cf. Langford\textsuperscript{8,9,10} (’05, ’08, ’09 ))

- consider numerator for deterministic policy:
  \[ p_+(a|x) = 1[a = h(x)] \]
- \[ E_+(Y) \propto \sum_i (y_i/p_-(a|x)) 1[a = h(x)] \equiv \sum_i w_i 1[a = h(x)] \]
- Note: \( 1[c = d] = 1 - 1[c \neq d] \)
- \[ \therefore E_+(Y) \propto \text{constant} - \sum_i w_i 1[a \neq h(x)] \]
- \[ \therefore \text{reduces policy optimization to (weighted) classification} \]

---

\textsuperscript{8}Langford & Zadrozny “Relating Reinforcement Learning Performance to Classification Performance” ICML 2005
\textsuperscript{9}Beygelzimer & Langford “The offset tree for learning with partial labels” (KDD 2009)
\textsuperscript{10}Tutorial on “Reductions” (including at ICML 2009)
Reduction w/ optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
Reduction w/optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
Reduction w/ optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
Reduction w/ optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
- normalize as $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}$
Reduction w/optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
- normalize as $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}$
- destroys lovely reduction
Reduction w/optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
- normalize as $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)] / p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)] / p_-(a_i|x_i)}$
- destroys lovely reduction
- simply\(^{11}\) $L(\lambda) = \sum_i (y_i - \lambda) 1[a_i \neq h(x_i)] / p_-(a_i|x_i)$

\(^{11}\)Suggestion by Dan Hsu
Reduction w/optimistic complication

- Prescription $\iff$ classification $L = \sum_i w_i 1[a_i \neq h(x_i)]$
- weight $w_i = y_i / p_-(a_i|x_i)$, inferred or RCT
- destroys measure by treating $p_-(a|x)$ differently than $1/p_-(a|x)$
- normalize as $\tilde{L} \equiv \frac{\sum_i y_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}{\sum_i 1[a_i \neq h(x_i)]/p_-(a_i|x_i)}$
- destroys lovely reduction
- simply $L(\lambda) = \sum_i (y_i - \lambda) 1[a_i \neq h(x_i)]/p_-(a_i|x_i)$
- hidden here is a 2nd parameter, in classification, \therefore harder search

POISE punchlines

- allows policy planning even with implicit logged exploration data\textsuperscript{12}

POISE punchlines

- allows policy planning even with implicit logged exploration data\textsuperscript{12}
- e.g., two hospital story

POISE punchlines

- allows policy planning even with implicit logged exploration data\textsuperscript{12}
- e.g., two hospital story
- “personalized medicine” is also a policy

POISE punchlines

- allows policy planning even with implicit logged exploration data\(^\text{12}\)
- e.g., two hospital story
- “personalized medicine” is also a policy
- abundant data available, under-explored IMHO

tangent: causality as told by an economist

different, related goal

- they think in terms of ATE/ITE instead of policy
tangent: causality as told by an economist

different, related goal

▶ they think in terms of ATE/ITE instead of policy
  ▶ ATE
different, related goal

- they think in terms of ATE/ITE instead of policy
  - ATE
    - \( \tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0) \)
different, related goal

- they think in terms of ATE/ITE instead of policy
  - ATE
    - $\tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0)$
  - CATE aka Individualized Treatment Effect (ITE)
different, related goal

- they think in terms of ATE/ITE instead of policy
  - ATE
    - $\tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0)$
  - CATE aka Individualized Treatment Effect (ITE)
    - $\tau(x) \equiv E_0(Y|a = 1, x) - E_0(Y|a = 0, x)$
tangent: causality as told by an economist

different, related goal

- they think in terms of ATE/ITE instead of policy
  
  - ATE
    
    \[ \tau \equiv E_0(Y|a = 1) - E_0(Y|a = 0) \equiv Q(a = 1) - Q(a = 0) \]
  
  - CATE aka Individualized Treatment Effect (ITE)
    
    \[ \tau(x) \equiv E_0(Y|a = 1, x) - E_0(Y|a = 0, x) \]
    
    \[ \equiv Q(a = 1, x) - Q(a = 0, x) \]
Q-note: “generalizing” Monte Carlo w/kernels

\[ MC: \quad E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i) \]
Q-note: “generalizing” Monte Carlo w/kernels

- **MC**: \( E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i) \)
- **K**: \( p \approx N^{-1} \sum_i K(x|x_i) \)
Q-note: “generalizing” Monte Carlo w/kernels

- **MC:** \( E_p(f) = \sum_x p(x) f(x) \approx N^{-1} \sum_{i \sim p} f(x_i) \)
- **K:** \( p \approx N^{-1} \sum_i K(x|x_i) \)
- \( \Rightarrow \sum_x p(x) f(x) \approx N^{-1} \sum_i \sum_x f(x) K(x|x_i) \)
Q-note: “generalizing” Monte Carlo w/kernels

- \( MC: \ E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i) \)
- \( K: \ p \approx N^{-1} \sum_i K(x|x_i) \)
- \( \Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i) \)
- \( K \) can be any normalized function, e.g., \( K(x|x_i) = \delta_{x,x_i} \), which yields \( MC. \)
Q-note: “generalizing” Monte Carlo w/kernels

- \( MC: \ E_p(f) = \sum_x p(x)f(x) \approx N^{-1} \sum_{i \sim p} f(x_i) \)
- \( K: \ p \approx N^{-1} \sum_i K(x|x_i) \)
- \( \Rightarrow \sum_x p(x)f(x) \approx N^{-1} \sum_i \sum_x f(x)K(x|x_i) \)
- \( K \) can be any normalized function, e.g., \( K(x|x_i) = \delta_{x,x_i} \), which yields \( MC \).
- \( \text{multivariate} \)
  \( E_p(f) \approx N^{-1} \sum_i \sum_{yax} f(y,a,x)K_1(y|y_i)K_2(a|a_i)K_3(x|x_i) \)
Q-note: application w/strata+matching, setup

Helps think about economists’ approach:

\[ Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y,a,x)}{p_-(a|x)p(x)} \]
Q-note: application w/strata+matching, setup

Helps think about economists’ approach:

\[ Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)} \]

\[ = \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x) \]
Q-note: application w/strata+matching, setup

Helps think about economists’ approach:

1. \( Q(a, x) \equiv E(Y \mid a, x) = \sum_y y p(y \mid a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a \mid x)p(x)} \)

2. \( = \frac{1}{p_-(a \mid x)p(x)} \sum_y y p_-(y, a, x) \)

3. stratify \( x \) using \( z(x) \) such that \( \bigcup z = X \), and \( \bigcap z, z' = \emptyset \)
Helps think about economists’ approach:

\[ Q(a, x) \equiv E(Y \mid a, x) = \sum_y y p(y \mid a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a \mid x) p(x)} \]

\[ = \frac{1}{p_-(a \mid x) p(x)} \sum_y y p_-(y, a, x) \]

stratify \( x \) using \( z(x) \) such that \( \cup z = X \), and \( \cap z, z' = \emptyset \)

\[ n(x) = \sum_i 1[z(x_i) = z(x)] \text{=number of points in} \ x \text{'s stratum} \]
Q-note: application w/strata+matching, setup

Helps think about economists’ approach:

- \( Q(a, x) \equiv E(Y|a, x) = \sum_y y p(y|a, x) = \sum_y y \frac{p_-(y,a,x)}{p_-(a|x)p(x)} \)
- \( = \frac{1}{p_-(a|x)p(x)} \sum_y y p_-(y, a, x) \)
- stratify \( x \) using \( z(x) \) such that \( \cup z = X \), and \( \cap z, z' = \emptyset \)
- \( n(x) = \sum_i 1[z(x_i) = z(x)] \)=number of points in \( x \)'s stratum
- \( \Omega(x) = \sum_{x'} 1[z(x') = z(x)] \)=area of \( x \)'s stratum
Q-note: application w/strata+matching, setup

Helps think about economists’ approach:

\[ Q(a, x) \equiv E(Y|a, x) = \sum_y yp(y|a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a|x)p(x)} \]

\[ = \frac{1}{p_-(a|x)p(x)} \sum_y yp_-(y, a, x) \]

stratify \( x \) using \( z(x) \) such that \( \cup z = X \), and \( \cap z, z' = \emptyset \)

\[ n(x) = \sum_i 1[z(x_i) = z(x)] = \text{number of points in } x \text{'s stratum} \]

\[ \Omega(x) = \sum_{x'} 1[z(x') = z(x)] = \text{area of } x \text{'s stratum} \]

\[ \therefore K_3(x|x_i) = \frac{1[z(x) = z(x_i)]}{\Omega(x)} \]
Helps think about economists’ approach:

- \( Q(a, x) \equiv E(Y | a, x) = \sum_y yp(y | a, x) = \sum_y y \frac{p_-(y, a, x)}{p_-(a | x)p(x)} \)
- \( = \frac{1}{p_-(a | x)p(x)} \sum_y yp_-(y, a, x) \)
- stratify \( x \) using \( z(x) \) such that \( \cup z = X \), and \( \cap z, z' = \emptyset \)
- \( n(x) = \sum_i 1[z(x_i) = z(x)] \) = number of points in \( x \)'s stratum
- \( \Omega(x) = \sum_{x'} 1[z(x') = z(x)] \) = area of \( x \)'s stratum
- \( \therefore K_3(x | x_i) = 1[z(x) = z(x_i)] / \Omega(x) \)
- as in \( MC \), \( K_1(y | y_i) = \delta_{y, y_i}, K_2(a | a_i) = \delta_{a, a_i} \)
Q-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]
Q-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]

\[ p(x) \approx (n(x)/N) \Omega(x)^{-1} \]
\[ Q\text{-note: application w/strata+matching, payoff} \]

- \[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]
- \[ p(x) \approx (n(x)/N) \Omega(x)^{-1} \]
- \[ \therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]
$Q$-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]

\[ p(x) \approx \left( \frac{n(x)}{N} \right) \Omega(x)^{-1} \]

\[ \therefore Q(a, x) \approx p_-(a | x)^{-1} n(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]
Q-note: application w/strata+matching, payoff

1. \[ \sum_{y} y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]
2. \[ p(x) \approx \left( \frac{n(x)}{N} \right) \Omega(x)^{-1} \]
3. \[ \therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]

“matching” means: choose each \( z \) to contain 1 positive example & 1 negative example,

4. \[ p_-(a|x) \approx 1/2, n(x) = 2 \]
Q-note: application w/strata+matching, payoff

- \[\sum_y yp_-(y, a, x) \approx N^{-1}\Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i\]
- \[p(x) \approx (n(x)/N)\Omega(x)^{-1}\]
- \[\therefore Q(a, x) \approx p_-(a|x)^{-1}n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i\]

“matching” means: choose each z to contain 1 positive example & 1 negative example,

- \[p_-(a|x) \approx 1/2, n(x) = 2\]
- \[\therefore \tau(a, x) = Q(a = 1, x) - Q(a = 0, x) = y_1(x) - y_0(x)\]
Q-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]

\[ p(x) \approx (n(x)/N) \Omega(x)^{-1} \]

\[ \therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]

“matching” means: choose each \( z \) to contain 1 positive example & 1 negative example,

\[ p_-(a|x) \approx 1/2, n(x) = 2 \]

\[ \therefore \tau(a, x) = Q(a = 1, x) - Q(a = 0, x) = y_1(x) - y_0(x) \]

\[ z \text{-generalizations: graphs, digraphs, k-NN, “matching”} \]
Q-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]

\[ p(x) \approx (n(x)/N) \Omega(x)^{-1} \]

\[ \therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i \]

“matching” means: choose each \( z \) to contain 1 positive example & 1 negative example,

\[ p_-(a|x) \approx 1/2, \ n(x) = 2 \]

\[ \therefore \tau(a, x) = Q(a = 1, x) - Q(a = 0, x) = y_1(x) - y_0(x) \]

\[ z \text{-generalizations: graphs, digraphs, k-NN, “matching”} \]

\[ K \text{-generalizations: continuous } a, \text{ any metric or similarity you like}, \ldots \]
Q-note: application w/strata + matching, payoff

\[
\sum_y yp_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i
\]

\[
p(x) \approx (n(x)/N) \Omega(x)^{-1}
\]

\[
\therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i=a, z(x_i)=z(x)} y_i
\]

“matching” means: choose each \( z \) to contain 1 positive example & 1 negative example,

\[
p_-(a|x) \approx 1/2, n(x) = 2
\]

\[
\therefore \tau(a, x) = Q(a = 1, x) - Q(a = 0, x) = y_1(x) - y_0(x)
\]

\[z\text{-generalizations: graphs, digraphs, k-NN, “matching”}\]

\[K\text{-generalizations: continuous } a, \text{ any metric or similarity you like,}\ldots\]
Q-note: application w/strata+matching, payoff

\[ \sum_y y p_-(y, a, x) \approx N^{-1} \Omega(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]

\[ p(x) \approx \frac{n(x)}{N} \Omega(x)^{-1} \]

\[ \therefore Q(a, x) \approx p_-(a|x)^{-1} n(x)^{-1} \sum_{a_i = a, z(x_i) = z(x)} y_i \]

“matching” means: choose each \( z \) to contain 1 positive example & 1 negative example,

\[ p_-(a|x) \approx 1/2, \quad n(x) = 2 \]

\[ \therefore \tau(a, x) = Q(a = 1, x) - Q(a = 0, x) = y_1(x) - y_0(x) \]

\[ z \text{-generalizations: graphs, digraphs, k-NN, “matching”} \]

\[ K \text{-generalizations: continuous } a, \text{ any metric or similarity you like, . . .} \]

IMHO underexplored
causality, as understood in marketing

- a/b testing and RCT

Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. Database Marketing, Springer New York, 2008
causality, as understood in marketing

- a/b testing and RCT
- yield optimization

Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. Database Marketing, Springer New York, 2008
causality, as understood in marketing

- a/b testing and RCT
- yield optimization
- Lorenz curve (vs ROC plots)

Figure 11: Blattberg, Robert C., Byung-Do Kim, and Scott A. Neslin. Database Marketing, Springer New York, 2008
unobserved confounders vs. “causality” modeling

truth: \( p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u) \)
unobserved confounders vs. “causality” modeling

- truth: \( p_\alpha(y, a, x, u) = p(y | a, x, u)p_\alpha(a | x, u)p(x, u) \)
- but: \( p_+(y, a, x, u) = p(y | a, x, u)p_-(a | x)p(x, u) \)
unobserved confounders vs. “causality” modeling

- truth: \[ p_\alpha(y, a, x, u) = p(y|a, x, u)p_\alpha(a|x, u)p(x, u) \]
- but: \[ p_+(y, a, x, u) = p(y|a, x, u)p_-(a|x)p(x, u) \]
- \[ E_+(Y) \equiv \sum_{yaxu} yp_+(yaxu) \approx N^{-1} \sum_{i \sim p_-} y_i p_+(a|x)/p_-(a|x, u) \]
unobserved confounders vs. “causality” modeling

- truth: \( p_{\alpha}(y, a, x, u) = p(y|a, x, u)p_{\alpha}(a|x, u)p(x, u) \)
- but: \( p_+(y, a, x, u) = p(y|a, x, u)p_-(a|x)p(x, u) \)
- \( E_+(Y) \equiv \sum_{yaxu} yp_+(yaxu) \approx \)
  \( N^{-1} \sum_{i \sim p_-} y_i p_+(a|x)/p_-(a|x, u) \)
- denominator can not be inferred, ignore at your peril
cautionary tale problem: Simpson’s paradox

- $a$: admissions (a=1: admitted, a=0: declined)
cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
- lawsuit (1973): $0.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = 0.35$
cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
- lawsuit (1973): $0.44 = p(a=1|x=0) > p(a=1|x=1) = 0.35$
- ‘resolved’ by Bickel (1975)$^{13}$ (See also Pearl$^{14}$)

---


cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
- lawsuit (1973): $0.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = 0.35$
- ‘resolved’ by Bickel (1975)$^{13}$ (See also Pearl$^{14}$)
- $u$: unobserved department they applied to

---


cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
- lawsuit (1973): $0.44 = p(a=1|x=0) > p(a=1|x=1) = 0.35$
- ‘resolved’ by Bickel (1975)$^{13}$ (See also Pearl$^{14}$)
- $u$: unobserved department they applied to
- $p(a|x) = \sum_{u=1}^{u=6} p(a|x,u)p(u|x)$

---


cautionary tale problem: Simpson’s paradox

- $a$: admissions ($a=1$: admitted, $a=0$: declined)
- $x$: gender ($x=1$: female, $x=0$: male)
- lawsuit (1973): $0.44 = p(a = 1|x = 0) > p(a = 1|x = 1) = 0.35$
- ‘resolved’ by Bickel (1975)\(^{13}\) (See also Pearl\(^{14}\))
- $u$: unobserved department they applied to
- $p(a|x) = \sum_{u=1}^{u=6} p(a|x, u)p(u|x)$
- e.g., gender-blind: $p(a|1) - p(a|0) = p(a|u) \cdot (p(u|1) - p(u|0))$


confounded approach: quasi-experiments + instruments

- Q: does engagement drive retention? (NYT, NFLX, . . . )
Q: does engagement drive retention? (NYT, NFLX, . . . )
   we don’t directly control engagement
confounded approach: quasi-experiments + instruments

- Q: does engagement drive retention? (NYT, NFLX, . . .)
  - we don’t directly control engagement
  - nonetheless useful since many things can influence it
Q: does engagement drive retention? (NYT, NFLX, ...)
   - we don’t directly control engagement
   - nonetheless useful since many things can influence it

Q: does serving in Vietnam war decrease earnings\(^\text{15}\)?

confounded approach: quasi-experiments + instruments

- Q: does engagement drive retention? (NYT, NFLX, . . .)
  - we don’t directly control engagement
  - nonetheless useful since many things can influence it
- Q: does serving in Vietnam war decrease earnings?  
  - US didn’t directly control serving in Vietnam, either

---


16 cf., George Bush, Donald Trump, Bill Clinton, Dick Cheney . . .
Q: does engagement drive retention? (NYT, NFLX, ...) 
- we don’t directly control engagement 
- nonetheless useful since many things can influence it

Q: does serving in Vietnam war decrease earnings? 
- US didn’t directly control serving in Vietnam, either

requires strong assumptions, including linear model

---

16 cf., George Bush, Donald Trump, Bill Clinton, Dick Cheney...
17 I thank Sinan Aral, MIT Sloan, for bringing this to my attention
IV: graphical model assumption

Figure 12: independence assumption
IV: graphical model assumption (sideways)

Figure 13: independence assumption
IV: review s/OLS/MOM/ ($E$ is empirical average)

- a endogenous
IV: review s/OLS/MOM/ (E is empirical average)

- a endogenous
  - e.g., $\exists u$ s.t. $p(y|a, x, u), p(a|x, u)$
IV: review s/OLS/MOM/ (E is empirical average)

- a endogenous
  - e.g., \( \exists u \text{ s.t. } p(y|a, x, u), p(a|x, u) \)
- linear ansatz: \( y = \beta^T a + \epsilon \)
IV: review s/OLS/MOM/ (E is empirical average)

- a endogenous
  - e.g., \( \exists u \text{ s.t. } p(y|a, x, u), p(a|x, u) \)
- linear ansatz: \( y = \beta^T a + \epsilon \)
- if a exogenous (e.g., OLS), use \( E[Y_{Aj}] = E[\beta^T A_{Aj}] + E[\epsilon A_j] \) (note that \( E[A_{Aj}A_{Ak}] \) gives square matrix; invert for \( \beta \))
IV: review s/OLS/MOM/ (E is empirical average)

- **a endogenous**
  - e.g., $\exists u \text{ s.t. } p(y|a,x,u), p(a|x,u)$
- linear ansatz: $y = \beta^T a + \epsilon$
- if a exogenous (e.g., OLS), use $E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j]$ (note that $E[A_j A_k]$ gives square matrix; invert for $\beta$)
- add instrument $x$ uncorrelated with $\epsilon$
IV: review s/OLS/MOM/ \( (E \text{ is empirical average}) \)

- a endogenous
  - e.g., \( \exists u \text{ s.t. } p(y|a,x,u), p(a|x,u) \)

- linear ansatz: \( y = \beta^T a + \epsilon \)
- if a exogenous (e.g., OLS), use \( E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j] \) (note that \( E[A_j A_k] \) gives square matrix; invert for \( \beta \))
- add instrument \( x \) uncorrelated with \( \epsilon \)
- \( E[YX_k] = E[\beta^T AX_k] + E[\epsilon]E[X_k] \)
IV: review s/OLS/MOM/ (E is empirical average)

- a endogenous
  - e.g., \( \exists u \ s.t. \ p(y|a, x, u), p(a|x, u) \)

- linear ansatz: \( y = \beta^T a + \epsilon \)
- if a exogenous (e.g., OLS), use \( E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j] \)
  (note that \( E[A_j A_k] \) gives square matrix; invert for \( \beta \))
- add instrument \( x \) uncorrelated with \( \epsilon \)
  - \( E[YX_k] = E[\beta^T AX_k] + E[\epsilon] E[X_k] \)
  - \( E[Y] = E[\beta^T A] + E[\epsilon] \) (from ansatz)
IV: review s/OLS/MOM/ (E is empirical average)

- a endogenous
  - e.g., \( \exists u \) s.t. \( p(y|a, x, u), p(a|x, u) \)
- linear ansatz: \( y = \beta^T a + \epsilon \)
- if \( a \) exogenous (e.g., OLS), use \( E[YA_j] = E[\beta^T AA_j] + E[\epsilon A_j] \) (note that \( E[A_j A_k] \) gives square matrix; invert for \( \beta \))
- add instrument \( x \) uncorrelated with \( \epsilon \)
- \( E[YX_k] = E[\beta^T AX_k] + E[\epsilon] E[X_k] \)
- \( E[Y] = E[\beta^T A] + E[\epsilon] \) (from ansatz)
- \( C(Y, X_k) = \beta^T C(A, X_k) \), not an “inversion” problem, requires “two stage regression”
IV: binary, binary case (aka “Wald estimator”)

\[ y = \beta a + \epsilon \]
IV: binary, binary case (aka “Wald estimator”)

- $y = \beta a + \epsilon$
- $E(Y|x) = \beta E(A|x) + E(\epsilon)$, evaluate at $x = \{0, 1\}$
IV: binary, binary case (aka “Wald estimator”)

- $y = \beta a + \epsilon$
- $E(Y|a) = \beta E(A) + E(\epsilon)$, evaluate at $a = \{0, 1\}$
- $\beta = (E(Y|a = 1) - E(Y|a = 0))/(E(A|a = 1) - E(A|a = 0))$. 
bandits: obligatory slide

Figure 14: almost all the talks I’ve gone to on bandits have this image
bandits

- wide applicability: humane clinical trials, targeting, ...
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
bandits

- wide applicability: humane clinical trials, targeting, . . .
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
- less useful for one-off, major decisions to be “interpreted”
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
- less useful for one-off, major decisions to be “interpreted”
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
- less useful for one-off, major decisions to be “interpreted”

examples

- $\epsilon$-greedy (no context, aka ‘vanilla’, aka ‘context-free’)
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
- less useful for one-off, major decisions to be “interpreted”

examples

- $\epsilon$-greedy (no context, aka ‘vanilla’, aka ‘context-free’)
- UCB1 (2002) (no context) + LinUCB (with context)
bandits

- wide applicability: humane clinical trials, targeting, ...
- replace meetings with code
- requires software engineering to replace decisions with, e.g., Javascript
- most useful if decisions or items get “stale” quickly
- less useful for one-off, major decisions to be “interpreted”

examples

- $\epsilon$-greedy (no context, aka ‘vanilla’, aka ‘context-free’) 
- UCB1 (2002) (no context) + LinUCB (with context) 
- Thompson Sampling (1933)$^{18,19,20}$ (general, with or without context)

---


$^{19}$AKA “probability matching”, “posterior sampling”

$^{20}$cf., “Bayesian Bandit Explorer” (link)
TS: connecting w/“generative causal modeling”

- WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$  
These 3 terms were treated by
WS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

- These 3 terms were treated by
  - response $p(y|a, x)$: avoid regression/inferring using importance sampling
TS: connecting w/“generative causal modeling”

- \( p(y, x, a) = p(y| x, a)p_\alpha(a|x)p(x) \)
- These 3 terms were treated by
  - response \( p(y|a, x) \): avoid regression/inferring using importance sampling
  - policy \( p_\alpha(a|x) \): optimize ours, infer theirs
WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

These 3 terms were treated by

- response $p(y|a, x)$: avoid regression/inferring using importance sampling
- policy $p_\alpha(a|x)$: optimize ours, infer theirs
- (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
TS: connecting w/“generative causal modeling”

- WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$
- These 3 terms were treated by
  - response $p(y|a, x)$: avoid regression/inferring using importance sampling
  - policy $p_\alpha(a|x)$: optimize ours, infer theirs
  - (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
  - prior $p(x)$: either avoid by importance sampling or estimate via kernel methods
WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

These 3 terms were treated by

- response $p(y|a, x)$: avoid regression/inferring using importance sampling
- policy $p_\alpha(a|x)$: optimize ours, infer theirs
- (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
- prior $p(x)$: either avoid by importance sampling or estimate via kernel methods

In the economics approach we focus on
TS: connecting w/“generative causal modeling” 0

- \[ p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x) \]
- These 3 terms were treated by
  - response \( p(y|a, x) \): avoid regression/inferring using importance sampling
  - policy \( p_\alpha(a|x) \): optimize ours, infer theirs
  - (NB: ours was deterministic: \( p(a|x) = 1[a = h(x)] \))
  - prior \( p(x) \): either avoid by importance sampling or estimate via kernel methods

- In the economics approach we focus on
  - \( \tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots) \) “treatment effect”
WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

These 3 terms were treated by

- **response** $p(y|a, x)$: avoid regression/inferring using importance sampling
- **policy** $p_\alpha(a|x)$: optimize ours, infer theirs
- (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
- **prior** $p(x)$: either avoid by importance sampling or estimate via kernel methods

In the economics approach we focus on

- $\tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots)$ “treatment effect”
- where $Q(a, \ldots) = \sum_y yp(y|\ldots)$
WAS \( p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x) \)

These 3 terms were treated by

- response \( p(y|a, x) \): avoid regression/inferring using importance sampling
- policy \( p_\alpha(a|x) \): optimize ours, infer theirs
- (NB: ours was deterministic: \( p(a|x) = 1[a = h(x)] \))
- prior \( p(x) \): either avoid by importance sampling or estimate via kernel methods

In the economics approach we focus on

- \( \tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots) \) “treatment effect”
- where \( Q(a, \ldots) = \sum_y yp(y|\ldots) \)
WAS $p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x)$

These 3 terms were treated by

- response $p(y|a, x)$: avoid regression/inferring using importance sampling
- policy $p_\alpha(a|x)$: optimize ours, infer theirs
- (NB: ours was deterministic: $p(a|x) = 1[a = h(x)]$)
- prior $p(x)$: either avoid by importance sampling or estimate via kernel methods

In the economics approach we focus on

- $\tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots)$ “treatment effect”
- where $Q(a, \ldots) = \sum_y yp(y|\ldots)$

In Thompson sampling we will generate 1 datum at a time, by

- asserting a parameterized generative model for $p(y|a, x, \theta)$
TS: connecting with “generative causal modeling”

- \( p(y, x, a) = p(y|x, a)p_\alpha(a|x)p(x) \)

- These 3 terms were treated by
  - response \( p(y|a, x) \): avoid regression/inferring using importance sampling
  - policy \( p_\alpha(a|x) \): optimize ours, infer theirs
  - (NB: ours was deterministic: \( p(a|x) = 1[a = h(x)] \))
  - prior \( p(x) \): either avoid by importance sampling or estimate via kernel methods

- In the economics approach we focus on
  - \( \tau(\ldots) \equiv Q(a = 1, \ldots) - Q(a = 0, \ldots) \) “treatment effect”
  - where \( Q(a, \ldots) = \sum_y yp(y|\ldots) \)

In Thompson sampling we will generate 1 datum at a time, by

- asserting a parameterized generative model for \( p(y|a, x, \theta) \)
- using a deterministic but averaged policy
TS: connecting w/“generative causal modeling”

- model true world response function $p(y|a,x)$ parametrically as $p(y|a,x,\theta^*)$
model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$
(i.e., $\theta^*$ is the true value of the parameter)\textsuperscript{21}

\textsuperscript{21}Note that $\theta$ is a vector, with components for each action.
model true world response function $p(y|a,x)$ parametrically as $p(y|a,x,\theta^*)$

(i.e., $\theta^*$ is the true value of the parameter)\(^{21}\)

if you knew $\theta$:

\(^{21}\)Note that $\theta$ is a vector, with components for each action.
model true world response function $p(y|a,x)$ parametrically as $p(y|a,x,\theta^*)$

(i.e., $\theta^*$ is the true value of the parameter)\[^{21}\]

if you knew $\theta$:

- could compute $Q(a,x,\theta) \equiv \sum_y yp(y|x,a,\theta^*)$ directly

\[^{21}\text{Note that } \theta \text{ is a vector, with components for each action.}\]
TS: connecting w/“generative causal modeling” 1

- model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$
- (i.e., $\theta^*$ is the true value of the parameter)$^{21}$
- if you knew $\theta$:
  - could compute $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
  - then choose $h(x; \theta) = \arg\max_a Q(a, x, \theta)$

$^{21}$Note that $\theta$ is a vector, with components for each action.
model true world response function \( p(y|a, x) \) parametrically as \( p(y|a, x, \theta^*) \)

(i.e., \( \theta^* \) is the true value of the parameter)

if you knew \( \theta \):

- could compute \( Q(a, x, \theta) \equiv \sum_y yp(y|x, a, \theta^*) \) directly
- then choose \( h(x; \theta) = \text{argmax}_a Q(a, x, \theta) \)
- inducing policy \( p(a|x, \theta) = 1[a = h(x; \theta) = \text{argmax}_a Q(a, x, \theta)] \)

\(^{21}\)Note that \( \theta \) is a vector, with components for each action.
model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$ (i.e., $\theta^*$ is the true value of the parameter)\(^{21}\)

if you knew $\theta$:

- could compute $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
- then choose $h(x; \theta) = \arg\max_a Q(a, x, \theta)$
- inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \arg\max_a Q(a, x, \theta)]$

idea: use prior data $D = \{y, a, x\}_1^t$ to define non-deterministic policy:

\(^{21}\)Note that $\theta$ is a vector, with components for each action.
model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$ (i.e., $\theta^*$ is the true value of the parameter)\textsuperscript{21}

if you knew $\theta$:

- could compute $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
- then choose $h(x; \theta) = \text{argmax}_a Q(a, x, \theta)$
- inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \text{argmax}_a Q(a, x, \theta)]$

idea: use prior data $D = \{y, a, x\}_{t=1}^T$ to define non-deterministic policy:

- $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$

\textsuperscript{21}Note that $\theta$ is a vector, with components for each action.
TS: connecting w/ “generative causal modeling”

- model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$
- (i.e., $\theta^*$ is the true value of the parameter)
- if you knew $\theta$:
  - could compute $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
  - then choose $h(x; \theta) = \arg\max_a Q(a, x, \theta)$
  - inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \arg\max_a Q(a, x, \theta)]$
- idea: use prior data $D = \{y, a, x\}^t_1$ to define non-deterministic policy:
  - $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  - $p(a|x) = \int d\theta 1[a = \arg\max_{a'} Q(a', x, \theta)]p(\theta|D)$

---

\( ^{21}\)Note that $\theta$ is a vector, with components for each action.
model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$
(\text{i.e., } \theta^* \text{ is the true value of the parameter})$^{21}$
if you knew $\theta$:
\begin{itemize}
  \item could compute $Q(a, x, \theta) = \sum_y y p(y|x, a, \theta^*)$ directly
  \item then choose $h(x; \theta) = \text{argmax}_a Q(a, x, \theta)$
  \item inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \text{argmax}_a Q(a, x, \theta)]$
\end{itemize}

idea: use prior data $D = \{y, a, x\}_1^t$ to define \textit{non-deterministic} policy:
\begin{itemize}
  \item $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
  \item $p(a|x) = \int d\theta 1[a = \text{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)$
\end{itemize}

\vspace{1cm}
$^{21}$Note that $\theta$ is a vector, with components for each action.
model true world response function \( p(y|a, x) \) parametrically as 
\( p(y|a, x, \theta^*) \)

(i.e., \( \theta^* \) is the true value of the parameter)\(^{21}\)

if you knew \( \theta \):

- could compute \( Q(a, x, \theta) \equiv \sum_y yp(y|x, a, \theta^*) \) directly
- then choose \( h(x; \theta) = \text{argmax}_a Q(a, x, \theta) \)
- inducing policy \( p(a|x, \theta) = 1[a = h(x; \theta) = \text{argmax}_a Q(a, x, \theta)] \)

idea: use prior data \( D = \{y, a, x\}_1^t \) to define non-deterministic policy:

\[
p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)
\]

\[
p(a|x) = \int d\theta 1[a = \text{argmax}_{a'} Q(a', x, \theta)]p(\theta|D)
\]

hold up:

- Q1: what's \( p(\theta|D) \)?

\(^{21}\)Note that \( \theta \) is a vector, with components for each action.
model true world response function $p(y|a, x)$ parametrically as $p(y|a, x, \theta^*)$

(i.e., $\theta^*$ is the true value of the parameter)

if you knew $\theta$:

- could compute $Q(a, x, \theta) \equiv \sum_y y p(y|x, a, \theta^*)$ directly
- then choose $h(x; \theta) = \arg\max_a Q(a, x, \theta)$
- inducing policy $p(a|x, \theta) = 1[a = h(x; \theta) = \arg\max_a Q(a, x, \theta)]$

idea: use prior data $D = \{y, a, x\}_1^t$ to define non-deterministic policy:

- $p(a|x) = \int d\theta p(a|x, \theta)p(\theta|D)$
- $p(a|x) = \int d\theta 1[a = \arg\max_{a'} Q(a', x, \theta)]p(\theta|D)$

hold up:

- Q1: what’s $p(\theta|D)$?
- Q2: how am I going to evaluate this integral?

\footnote{Note that $\theta$ is a vector, with components for each action.}
TS: connecting w/ “generative causal modeling”

- Q1: what’s $p(\theta|D)$?
Q1: what’s \( p(\theta | D) \)?

Q2: how am I going to evaluate this integral?
Q1: what’s $p(\theta|D)$?
Q2: how am I going to evaluate this integral?
A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$. 
Q1: what’s $p(\theta|D)$?
Q2: how am I going to evaluate this integral?
A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)
Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

$p(\theta|D) = p(\theta|\{y\}^t_1, \{a\}^t_1, \{x\}^t_1, \alpha)$
Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

$p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$

$\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$
Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

- $p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$
- $\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$
- $= p(\theta|\alpha)\prod_t p(y_t|a_t, x_t, \theta)$
Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

$p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$

$\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$

$= p(\theta|\alpha)\prod_t p(y_t|a_t, x_t, \theta)$

**warning 1:** sometimes people write “$p(D|\theta)$” but we don’t need $p(a|\theta)$ or $p(x|\theta)$ here
Q1: what’s \( p(\theta|D) \)?

Q2: how am I going to evaluate this integral?

A1: \( p(\theta|D) \) definable by choosing prior \( p(\theta|\alpha) \) and likelihood on \( y \) given by the (modeled, parameterized) response \( p(y|a, x, \theta) \).

(now you’re not only generative, you’re Bayesian.)

\[
p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)
\]
\[
\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)
\]
\[
= p(\theta|\alpha)\prod_tp(y_t|a_t, x_t, \theta)
\]

**warning 1**: sometimes people write “\( p(D|\theta) \)” but we don’t need \( p(a|\theta) \) or \( p(x|\theta) \) here

**warning 2**: don’t need historical record of \( \theta_t \).
Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

$p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$

$\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$

$= p(\theta|\alpha)\prod_t p(y_t|a_t, x_t, \theta)$

**warning 1**: sometimes people write “$p(D|\theta)$” but we don’t need $p(a|\theta)$ or $p(x|\theta)$ here

**warning 2**: don’t need historical record of $\theta_t$.

(we used Bayes rule, but only in $\theta$ and $y$.)
Q1: what’s $p(\theta|D)$?
Q2: how am I going to evaluate this integral?
A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a, x, \theta)$.

(now you’re not only generative, you’re Bayesian.)

\[
p(\theta|D) = p(\theta|\{y\}^t_1, \{a\}^t_1, \{x\}^t_1, \alpha) \\
\propto p(\{y\}^t_1|\{a\}^t_1, \{x\}^t_1, \theta)p(\theta|\alpha) \\
= p(\theta|\alpha)\prod_t p(y_t|a_t, x_t, \theta)
\]

warning 1: sometimes people write “$p(D|\theta)$” but we don’t need $p(a|\theta)$ or $p(x|\theta)$ here

warning 2: don’t need historical record of $\theta_t$.

(we used Bayes rule, but only in $\theta$ and $y$.)

A2: evaluate integral by $N=1$ Monte Carlo\footnote{\text{AFAIK it is open research area what happens when you replace $N=1$ with $N= N_0 t^{-\nu}$.}}
Q1: what’s $p(\theta|D)$?
Q2: how am I going to evaluate this integral?
A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a,x,\theta)$.

(now you’re not only generative, you’re Bayesian.)

$p(\theta|D) = p(\theta|\{y\}_1^t, \{a\}_1^t, \{x\}_1^t, \alpha)$

$\propto p(\{y\}_1^t|\{a\}_1^t, \{x\}_1^t, \theta)p(\theta|\alpha)$

$= p(\theta|\alpha)\prod_t p(y_t|a_t, x_t, \theta)$

warning 1: sometimes people write “$p(D|\theta)$” but we don’t need $p(a|\theta)$ or $p(x|\theta)$ here

warning 2: don’t need historical record of $\theta_t$.

(we used Bayes rule, but only in $\theta$ and $y$.)

A2: evaluate integral by $N = 1$ Monte Carlo\(^{22}\)

take 1 sample “$\theta_t$” of $\theta$ from $p(\theta|D)$

\(^{22}\)AFAIK it is open research area what happens when you replace $N = 1$ with $N = N_0 t^{-\nu}$. 

Q1: what’s $p(\theta|D)$?

Q2: how am I going to evaluate this integral?

A1: $p(\theta|D)$ definable by choosing prior $p(\theta|\alpha)$ and likelihood on $y$ given by the (modeled, parameterized) response $p(y|a,x,\theta)$.

- (now you’re not only generative, you’re Bayesian.)
- $p(\theta|D) = p(\theta|\{y\}^t_1,\{a\}^t_1,\{x\}^t_1,\alpha)$
- $\propto p(\{y\}^t_1|\{a\}^t_1,\{x\}^t_1,\theta)p(\theta|\alpha)$
- $= p(\theta|\alpha)\prod_tp(y_t|a_t,x_t,\theta)$

**warning 1:** sometimes people write “$p(D|\theta)$” but we don’t need $p(a|\theta)$ or $p(x|\theta)$ here

**warning 2:** don’t need historical record of $\theta_t$.

- (we used Bayes rule, but only in $\theta$ and $y$.)

A2: evaluate integral by $N=1$ Monte Carlo$^{22}$

- take 1 sample “$\theta_t$” of $\theta$ from $p(\theta|D)$
- $a_t = h(x_t;\theta_t) = \arg\max_a Q(a,x,\theta_t)$

$^{22}$AFAIK it is open research area what happens when you replace $N = 1$ with $N = N_0 t^{-\nu}$. 
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\},$
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\}$,
- no context $x$, 
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\}$,
- no context $x$,
- Bernoulli (coin flipping), keep track of
That sounds hard.

No, just general. Let’s do toy case:

- \( y \in \{0, 1\} \),
- no context \( x \),
- Bernoulli (coin flipping), keep track of
  - \( S_a \equiv \text{number of successes flipping coin } a \)
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\}$,
- no context $x$,
- Bernoulli (coin flipping), keep track of
  - $S_a \equiv$ number of successes flipping coin $a$
  - $F_a \equiv$ number of failures flipping coin $a$
That sounds hard.

No, just general. Let’s do toy case:

- \( y \in \{0, 1\} \),
- no context \( x \),
- Bernoulli (coin flipping), keep track of
  - \( S_a \equiv \) number of successes flipping coin \( a \)
  - \( F_a \equiv \) number of failures flipping coin \( a \)
That sounds hard.

No, just general. Let’s do toy case:

- \( y \in \{0, 1\} \),
- no context \( x \),
- Bernoulli (coin flipping), keep track of
  - \( S_a \equiv \text{number of successes flipping coin } a \)
  - \( F_a \equiv \text{number of failures flipping coin } a \)

Then

\[
p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)
\]
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\}$,
- no context $x$,
- Bernoulli (coin flipping), keep track of
  - $S_a \equiv$ number of successes flipping coin $a$
  - $F_a \equiv$ number of failures flipping coin $a$

Then

- $p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$
- $= \left(\prod_a \theta_a^{\alpha-1} (1 - \theta_a)^{\beta-1}\right) \left(\prod_{t,a} \theta_{a_t}^{y_t} (1 - \theta_{a_t})^{1-y_t}\right)$
That sounds hard.

No, just general. Let’s do toy case:

- $y \in \{0, 1\}$,
- no context $x$,
- Bernoulli (coin flipping), keep track of
  - $S_a \equiv$ number of successes flipping coin $a$
  - $F_a \equiv$ number of failures flipping coin $a$

Then

$$p(\theta|D) \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta)$$
$$= \left(\prod_a \theta_a^{\alpha-1}(1 - \theta_a)^{\beta-1}\right) \left(\prod_{t,a_t} \theta_{a_t}^{y_t} (1 - \theta_{a_t})^{1-y_t}\right)$$
$$= \prod_a \theta^{\alpha+S_a-1}(1 - \theta_a)^{\beta+F_a-1}$$
That sounds hard.

No, just general. Let’s do toy case:

- \( y \in \{0, 1\} \),
- no context \( x \),
- Bernoulli (coin flipping), keep track of
  - \( S_a \equiv \text{number of successes flipping coin } a \)
  - \( F_a \equiv \text{number of failures flipping coin } a \)

Then

\[
\begin{align*}
  p(\theta|D) & \propto p(\theta|\alpha) \prod_t p(y_t|a_t, \theta) \\
  & = \left( \prod_a \theta_a^{\alpha-1}(1 - \theta_a)^{\beta-1} \right) \left( \prod_t, a_t \theta_a^{y_t}(1 - \theta_a)^{1-y_t} \right) \\
  & = \prod_a \theta_\alpha^{S_a-1}(1 - \theta_a)^{F_a-1} \\
  \therefore \theta_a & \sim \text{Beta}(\alpha + S_a, \beta + F_a)
\end{align*}
\]
An Empirical Evaluation of Thompson Sampling

Olivier Chapelle
Yahoo! Research
Santa Clara, CA
chap@yahoo-inc.com

Lihong Li
Yahoo! Research
Santa Clara, CA
lihong@yahoo-inc.com

Figure 15: Chaleppe and Li 2011
In the realizable case, the reward is a stochastic function of the action, context and the unknown, true parameter $\theta^*$. Ideally, we would like to choose the action maximizing the expected reward, \[ \max_a \mathbb{E}(r|a, x, \theta^*). \]

Of course, $\theta^*$ is unknown. If we are just interested in maximizing the immediate reward (exploitation), then one should choose the action that maximizes \[ \mathbb{E}(r|a, x) = \int \mathbb{E}(r|a, x, \theta) P(\theta|D) d\theta. \]

But in an exploration / exploitation setting, the probability matching heuristic consists in randomly selecting an action $a$ according to its probability of being optimal. That is, action $a$ is chosen with probability

\[ \int \mathbb{I} \left[ \mathbb{E}(r|a, x, \theta) = \max_{a'} \mathbb{E}(r|a', x, \theta) \right] P(\theta|D) d\theta, \]

where $\mathbb{I}$ is the indicator function. Note that the integral does not have to be computed explicitly; it suffices to draw a random parameter $\theta$ at each round as explained in Algorithm 1. Implementation of the algorithm is thus efficient and straightforward in most applications.

Figure 16: from Chaleppe and Li 2011
Algorithm 1 Thompson sampling

\[ D = \emptyset \]
\[ \text{for } t = 1, \ldots, T \text{ do} \]
\[-\text{Receive context } x_t \]
\[-\text{Draw } \theta^t \text{ according to } P(\theta|D) \]
\[-\text{Select } a_t = \arg \max_a \mathbb{E}_r(r|x_t, a, \theta^t) \]
\[-\text{Observe reward } r_t \]
\[-D = D \cup (x_t, a_t, r_t) \]
\[\text{end for}\]
Algorithm 2 Thompson sampling for the Bernoulli bandit

Require: $\alpha, \beta$ prior parameters of a Beta distribution
$S_i = 0, F_i = 0, \forall i$. \{Success and failure counters\}

for $t = 1, \ldots, T$ do
  for $i = 1, \ldots, K$ do
    Draw $\theta_i$ according to $\text{Beta}(S_i + \alpha, F_i + \beta)$.
  end for
  Draw arm $\hat{i} = \arg \max_i \theta_i$ and observe reward $r$
  if $r = 1$ then
    $S_{\hat{i}} = S_{\hat{i}} + 1$
  else
    $F_{\hat{i}} = F_{\hat{i}} + 1$
  end if
end for

Figure 18: from Chaloppe and Li 2011
TS: Bernoulli bandit p-code (results)

Figure 19: from Chaleppe and Li 2011
Deterministic policy: \textbf{UCB1}.
Initialization: Play each machine once.
Loop:

- Play machine $j$ that maximizes $\bar{x}_j + \sqrt{\frac{2\ln n}{n_j}}$, where $\bar{x}_j$ is the average reward obtained from machine $j$, $n_j$ is the number of times machine $j$ has been played so far, and $n$ is the overall number of plays done so far.
Algorithm 3 Regularized logistic regression with batch updates

Require: Regularization parameter $\lambda > 0$.

$m_i = 0$, $q_i = \lambda$. {Each weight $w_i$ has an independent prior $\mathcal{N}(m_i, q_i^{-1})$}

for $t = 1, \ldots, T$ do

Get a new batch of training data $(x_j, y_j)$, $j = 1, \ldots, n$.

Find $w$ as the minimizer of:

$$
\frac{1}{2} \sum_{i=1}^{d} q_i (w_i - m_i)^2 + \sum_{j=1}^{n} \log(1 + \exp(-y_j w^T x_j)).
$$

$m_i = w_i$

$q_i = q_i + \sum_{j=1}^{n} x_{ij}^2 p_j (1 - p_j), \quad p_j = (1 + \exp(-w^T x_j))^{-1}$ \{Laplace approximation\}

end for

Figure 21: from Chaleppe and Li 2011
Algorithm 1 LinUCB with disjoint linear models.

0: Inputs: $\alpha \in \mathbb{R}_+$

1: for $t = 1, 2, 3, \ldots, T$ do

2: Observe features of all arms $a \in A_t$: $x_{t,a} \in \mathbb{R}^d$

3: for all $a \in A_t$ do

4: if $a$ is new then

5: \[ A_{\alpha} \leftarrow I_d \text{ (d-dimensional identity matrix)} \]

6: \[ b_{\alpha} \leftarrow 0_{d \times 1} \text{ (d-dimensional zero vector)} \]

7: end if

8: \[ \hat{\theta}_{\alpha} \leftarrow A_{\alpha}^{-1} b_{\alpha} \]

9: \[ p_{t,\alpha} \leftarrow \hat{\theta}_{\alpha}^T x_{t,a} + \alpha \sqrt{x_{t,a}^T A_{\alpha}^{-1} x_{t,a}} \]

10: end for

11: Choose arm $a_t = \arg \max_{a \in A_t} p_{t,a}$ with ties broken arbitrarily, and observe a real-valued payoff $r_t$

12: \[ A_{a_t} \leftarrow A_{a_t} + x_{t,a_t} x_{t,a_t}^T \]

13: \[ b_{a_t} \leftarrow b_{a_t} + r_t x_{t,a_t} \]

14: end for

Figure 22: LinUCB
Table 2: CTR regrets on the display advertising data.

<table>
<thead>
<tr>
<th>Method Parameter</th>
<th>TS 0.25</th>
<th>TS 0.5</th>
<th>TS 1</th>
<th>LinUCB 0.5</th>
<th>LinUCB 1</th>
<th>LinUCB 2</th>
<th>(\varepsilon)-greedy 0.005</th>
<th>(\varepsilon)-greedy 0.01</th>
<th>(\varepsilon)-greedy 0.02</th>
<th>Exploit</th>
<th>Randor</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regret (%)</td>
<td>4.45</td>
<td>3.72</td>
<td>3.81</td>
<td>4.99</td>
<td>4.22</td>
<td>4.14</td>
<td>5.05</td>
<td>4.98</td>
<td>5.22</td>
<td>5.00</td>
<td>31.95</td>
</tr>
</tbody>
</table>

Figure 4: CTR regret over the 4 days test period for 3 algorithms: Thompson sampling with \(\alpha = 0.5\), LinUCB with \(\alpha = 2\), Exploit-only. The regret in the first hour is large, around 0.3, because the algorithms predict randomly (no initial model provided).

Figure 23: from Chalapte and Li 2011
THEOREM 2. Assume that \( I(\theta, \lambda) \) satisfies (1.6) and (1.7) and that \( \Theta \) satisfies (1.9). Fix \( j \in \{1, \ldots, k\} \), and define \( \Theta_j \) and \( \Theta_j^* \) by (2.1). Let \( \varphi \) be any rule such that for every \( \theta \in \Theta_j^* \), as \( n \to \infty \)

\[
\sum_{i \neq j} E_\theta T_n(i) = o(n^a) \quad \text{for every } a > 0, \tag{2.2}
\]

where \( T_n(i) \), defined in (1.2), is the number of times that the rule \( \varphi \) samples from \( \Pi_i \) up to stage \( n \). Then for every \( \theta \in \Theta_j \) and every \( \epsilon > 0 \),

\[
\lim_{n \to \infty} P_\theta \left\{ T_n(j) \geq (1 - \epsilon)(\log n) / I(\theta_j, \theta^*) \right\} = 1, \tag{2.3}
\]

where \( \theta^* \) is defined in (1.4), and hence

\[
\liminf_{n \to \infty} E_\theta T_n(j) / \log n \geq 1 / I(\theta_j, \theta^*). 
\]
Thompson sampling (1933) and optimality (2013)

**Theorem 2.** For any instance $\Theta = \{\mu_1, \ldots, \mu_N\}$ of Bernoulli MAB,

$$R(T, \Theta) \leq (1 + \epsilon) \sum_{i \neq I^*} \frac{\ln(T) \Delta_i}{\text{KL}(\mu_i, \mu^*)} + O(N/\epsilon^2)$$

Recall that we have $\lim_{T \to \infty} \frac{R(T, \Theta)}{\ln(T)} \geq \sum_{i \neq I^*} \frac{\Delta_i}{\text{KL}(\mu_i, \mu^*)}$. Above theorem says that Thompson Sampling matches this lower bound. We also have the following problem independent regret bound for this algorithm.

**Theorem 3.** For all $\Theta$,

$$R(T) = \max_{\Theta} R(T, \Theta) \leq O(\sqrt{NT \log T} + N)$$

For proofs of above theorems, refer to [2].

Figure 25: TS result

other ‘Causalities’: structure learning

Figure 26: from heckerman 1995

other ‘Causalities’: potential outcomes

- model distribution of $p(y_i(1), y_i(0), a_i, x_i)$
other ‘Causalities’: potential outcomes

- model distribution of \( p(y_i(1), y_i(0), a_i, x_i) \)
- “action” replaced by “observed outcome”
other ‘Causalities’: potential outcomes

- model distribution of $p(y_i(1), y_i(0), a_i, x_i)$
- “action” replaced by “observed outcome”
- aka Neyman-Rubin causal model: Neyman (’23); Rubin (’74)
other ‘Causalities’: potential outcomes

- model distribution of $p(y_i(1), y_i(0), a_i, x_i)$
- “action” replaced by “observed outcome”
- aka Neyman-Rubin causal model: Neyman (’23); Rubin (’74)
- see Morgan + Winship\textsuperscript{24} for connections between frameworks

Lecture 4: descriptive modeling © NYT
what does kmeans mean?
review: (latent) inference and clustering

- what does kmeans mean?
  - given $x_i \in \mathbb{R}^D$
what does kmeans mean?

- given $x_i \in \mathbb{R}^D$
- given $d: \mathbb{R}^D \rightarrow \mathbb{R}^1$
what does kmeans mean?

- given \( x_i \in R^D \)
- given \( d : R^D \rightarrow R^1 \)
- assign \( z_i \)
review: (latent) inference and clustering

- what does kmeans mean?
  - given $x_i \in \mathbb{R}^D$
  - given $d : \mathbb{R}^D \rightarrow \mathbb{R}^1$
  - assign $z_i$

- generative modeling gives meaning
what does kmeans mean?
- given $x_i \in \mathbb{R}^D$
- given $d : \mathbb{R}^D \rightarrow \mathbb{R}^1$
- assign $z_i$

generative modeling gives meaning
- given $p(x|z, \theta)$
what does kmeans mean?
  - given $x_i \in \mathbb{R}^D$
  - given $d: \mathbb{R}^D \rightarrow \mathbb{R}^1$
  - assign $z_i$

generative modeling gives meaning
  - given $p(x|z, \theta)$
  - maximize $p(x|\theta)$
what does kmeans mean?

- given $x_i \in R^D$
- given $d : R^D \rightarrow R^1$
- assign $z_i$

generative modeling gives meaning

- given $p(x|z, \theta)$
- maximize $p(x|\theta)$
- output assignment $p(z|x, \theta)$
actual math

- define $P \equiv p(x, z|\theta)$
- define $P \equiv p(x, z|\theta)$
- log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$
  (cf. importance sampling)
actual math

- define $P \equiv p(x, z|\theta)$
- log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$
  (cf. importance sampling)
- Jensen’s:
  $L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$
define \( P \equiv p(x, z|\theta) \)

log-likelihood \( L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q \)
(cf. importance sampling)

Jensen’s:
\[
L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}
\]

analogy to free energy in physics
define \( P \equiv p(x, z|\theta) \)

log-likelihood \( L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q \)  
(cf. importance sampling)

Jensen’s:
\[
L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}
\]

- analogy to free energy in physics

alternate optimization on \( \theta \) and on \( q \)
define $P \equiv p(x, z|\theta)$
log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$
(cf. importance sampling)

Jensen’s:
$L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$

analog to free energy in physics

alternate optimization on $\theta$ and on $q$

NB: $q$ step gives $q(z) = p(z|x, \theta)$
define $P \equiv p(x, z|\theta)$

log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P/q$
(cf. importance sampling)

Jensen’s:
$L \geq \tilde{L} \equiv E_q \log P/q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$

- analogy to free energy in physics

alternate optimization on $\theta$ and on $q$

- NB: $q$ step gives $q(z) = p(z|x, \theta)$
- NB: log $P$ convenient for independent examples w/ exponential families
define $P \equiv p(x, z|\theta)$

- log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log \mathbb{E}_q P / q$
  (cf. importance sampling)

- Jensen’s:
  
  $$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$$

  - analogy to free energy in physics

- alternate optimization on $\theta$ and on $q$

  - NB: $q$ step gives $q(z) = p(z|x, \theta)$
  - NB: log $P$ convenient for independent examples w/ exponential families

  - e.g., GMMs: $\mu_k \leftarrow E[x|z]$ and $\sigma^2_k \leftarrow E[(x - \mu)^2|z]$ are sufficient statistics
define $P \equiv p(x, z|\theta)$

log-likelihood $L \equiv \log p(x|\theta) = \log \sum_z P = \log E_q P / q$

cf. importance sampling

Jensen’s:
$L \geq \tilde{L} \equiv E_q \log P / q = E_q \log P + H[q] = -(U - H) = -\mathcal{F}$

analogy to free energy in physics

alternate optimization on $\theta$ and on $q$

NB: $q$ step gives $q(z) = p(z|x, \theta)$

NB: log $P$ convenient for independent examples w/ exponential families

e.g., GMMs: $\mu_k \leftarrow E[x|z]$ and $\sigma_k^2 \leftarrow E[(x - \mu)^2|z]$ are sufficient statistics

e.g., LDAs: word counts are sufficient statistics
Energy $U$ (to be minimized):

\[ -U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z \]
Energy $U$ (to be minimized):

\[ -U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z \]

\[ -U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i | z_i) \]
Energy $U$ (to be minimized):

- $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i | z_i)$
- $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i | z_i)$
Energy $U$ (to be minimized):

- $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i | z_i)$
- $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i | z_i)$
- define $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
Energy $U$ (to be minimized):

- $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- Define $r_{ik} = \sum_z q_i(z) 1[z_i = k]$
- $-U_x = \sum_i r_{ik} \log p(x_i|k)$. 

---

tangent: more math on GMMs, part 1
Energy $U$ (to be minimized):

$-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$

$-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i | z_i)$

$= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i | z_i)$

define $r_{ik} = \sum_z q_i(z) 1[z_i = k]$

$-U_x = \sum_i r_{ik} \log p(x_i | k)$.

Gaussian$^{25}$

$\Rightarrow -U_x = \sum_i r_{ik} \left(-\frac{1}{2} (x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

$^{25}$math is simpler if you work with $\lambda_k \equiv \sigma^{-2}$
Energy $U$ (to be minimized):

- $-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$
- $-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$
- $= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$
- define $r_{ik} = \sum_z q_i(z)1[z_i = k]$
- $-U_x = \sum_i r_{ik} \log p(x_i|k)$.
- Gaussian\(^{25}\)

$\Rightarrow -U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi\right)$

\(^{25}\)math is simpler if you work with $\lambda_k \equiv \sigma^{-2}$
Energy $U$ (to be minimized):

$-U \equiv E_q \log P = \sum_z \sum_i q_i(z) \log P(x_i, z_i) \equiv U_x + U_z$

$-U_x \equiv \sum_z \sum_i q_i(z) \log p(x_i|z_i)$

$= \sum_i \sum_z q_i(z) \sum_k 1[z_i = k] \log p(x_i|z_i)$

define $r_{ik} = \sum_z q_i(z) 1[z_i = k]$

$-U_x = \sum_i r_{ik} \log p(x_i|k)$.

Gaussian$^{25}$

$\Rightarrow -U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right)$

simple to minimize for parameters $\vartheta = \{\mu_k, \lambda_k\}$

---

$^{25}$math is simpler if you work with $\lambda_k \equiv \sigma^{-2}$
- \( U_x = \sum_i r_{ik} \left( -\frac{1}{2} (x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right) \)
\[ -U_x = \sum_i r_{ik} \left( -\frac{1}{2} (x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi \right) \]

\[ \mu_k \leftarrow E[x \mid k] \text{ solves } \sum_i r_{ik} = \sum_i r_{ik} x_i \]
- $U_x = \sum_i r_{ik} \left(-\frac{1}{2}(x_i - \mu_k)^2 \lambda_k + \frac{1}{2} \ln \lambda_k - \frac{1}{2} \ln 2\pi\right)$
- $\mu_k \leftarrow E[x|k]$ solves $\sum_i r_{ik} = \sum_i r_{ik} x_i$
- $\lambda_k \leftarrow E[(x - \mu)^2|k]$ solves $\sum_i r_{ik} \frac{1}{2}(x_i - \mu_k)^2 = \lambda_k^{-1} \sum_i r_{ik}$
tangent: Gaussians $\in$ exponential family$^{27}$

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
tangent: Gaussians ∈ exponential family

- as before, \(-U = \sum_i r_{ik} \log p(x_i|k)\)
- define \(p(x_i|k) = \exp (\eta(\theta) \cdot T(x) - A(\theta) + B(x))\)
tangent: Gaussians $\in$ exponential family

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp (\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case \(^\text{26}\),

\(^\text{26}\)Choosing $\eta(\theta) = \eta$ called ‘canonical form’
tangent: Gaussians $\in$ exponential family\textsuperscript{27}

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp (\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case\textsuperscript{26},
  - $T_1 = x$,

\textsuperscript{26}Choosing $\eta(\theta) = \eta$ called ‘canonical form’
tangent: Gaussians ∈ exponential family

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case,
  - $T_1 = x,$
  - $T_2 = x^2$

\footnote{Choosing $\eta(\theta) = \eta$ called `canonical form`}
tangent: Gaussians ∈ exponential family\textsuperscript{27}

- as before, \(-U = \sum_i r_{ik} \log p(x_i|k)\)
- define \(p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))\)
- e.g., Gaussian case \textsuperscript{26},
  - \(T_1 = x\),
  - \(T_2 = x^2\)
  - \(\eta_1 = \mu/\sigma^2 = \mu\lambda\)

\textsuperscript{26}Choosing \(\eta(\theta) = \eta\) called ‘canonical form’
tangent: Gaussians ∈ exponential family

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp (\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case:
  - $T_1 = x$, $\eta_1 = \mu/\sigma^2 = \mu \lambda$
  - $T_2 = x^2$, $\eta_2 = -\frac{1}{2} \lambda = -1/(2\sigma^2)$

---

26 Choosing $\eta(\theta) = \eta$ called ‘canonical form’
tangent: Gaussians ∈ exponential family\(^{27}\)

- as before, \(-U = \sum_i r_{ik} \log p(x_i|k)\)
- define \(p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))\)
- e.g., Gaussian case \(^{26}\),
  - \(T_1 = x\),
  - \(T_2 = x^2\)
  - \(\eta_1 = \mu/\sigma^2 = \mu \lambda\)
  - \(\eta_2 = -\frac{1}{2} \lambda = -1/(2\sigma^2)\)
  - \(A = \lambda \mu^2/2 - \frac{1}{2} \ln \lambda\)

\(^{26}\)Choosing \(\eta(\theta) = \eta\) called ‘canonical form’
tangent: Gaussians ∈ exponential family

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp(\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case

\begin{itemize}
  \item $T_1 = x$,
  \item $T_2 = x^2$
  \item $\eta_1 = \mu/\sigma^2 = \mu\lambda$
  \item $\eta_2 = -\frac{1}{2}\lambda = -1/(2\sigma^2)$
  \item $A = \lambda\mu^2/2 - \frac{1}{2}\ln\lambda$
  \item $\exp(B(x)) = (2\pi)^{-1/2}$
\end{itemize}

\footnote{Choosing $\eta(\theta) = \eta$ called ‘canonical form’}
tangent: Gaussians $\in$ exponential family$^{27}$

- as before, $-U = \sum_i r_{ik} \log p(x_i|k)$
- define $p(x_i|k) = \exp (\eta(\theta) \cdot T(x) - A(\theta) + B(x))$
- e.g., Gaussian case $^{26}$,
  - $T_1 = x$,
  - $T_2 = x^2$
  - $\eta_1 = \mu / \sigma^2 = \mu \lambda$
  - $\eta_2 = -\frac{1}{2} \lambda = -1/(2\sigma^2)$
  - $A = \lambda \mu^2 / 2 - \frac{1}{2} \ln \lambda$
  - $\exp(B(x)) = (2\pi)^{-1/2}$

- note that in a mixture model, there are separate $\eta$ (and thus $A(\eta)$) for each value of $z$

$^{26}$Choosing $\eta(\theta) = \eta$ called ‘canonical form’

$^{27}$NB: Gaussians $\in$ exponential family, GMM $\notin$ exponential family! (Thanks to Eszter Vértes for pointing out this error in earlier title.)
tangent: variational joy $\in$ exponential family

\[ as \text{ before, } -U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right) \]
tangent: variational joy $\in$ exponential family

- as before, $-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$

- $\eta_{k,\alpha}$ solves $\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$ (canonical)
as before, \(-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)\)

\(\eta_{k,\alpha}\) solves \(\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}\) (canonical)

\(\therefore \partial_{\eta_{k,\alpha}} A(\eta_k) \leftarrow E[T_{k,\alpha}|k]\) (canonical)
tangent: variational joy ∈ exponential family

- as before, $$-U = \sum_i r_{ik} \left( \eta_k^T T(x_i) - A(\eta_k) + B(x_i) \right)$$
- $$\eta_{k,\alpha}$$ solves $$\sum_i r_{ik} T_{k,\alpha}(x_i) = \frac{\partial A(\eta_k)}{\partial \eta_{k,\alpha}} \sum_i r_{ik}$$ (canonical)
- $$\therefore \partial_{\eta_{k,\alpha}} A(\eta_k) \leftarrow E[T_{k,\alpha}|k]$$ (canonical)
- nice connection w/physics, esp. mean field theory

---

clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
- large freedom to choose different optimization approaches
clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
- large freedom to choose different optimization approaches
  - e.g., hard clustering limit
clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
- large freedom to choose different optimization approaches
  - e.g., hard clustering limit
  - e.g., streaming solutions
clustering and inference: GMM/k-means case study

- generative model gives meaning and optimization
- large freedom to choose different optimization approaches
  - e.g., hard clustering limit
  - e.g., streaming solutions
  - e.g., stochastic gradient methods
general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
- e.g., some favorite applications:
general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
- e.g., some favorite applications:
  - hmm
general framework: E+M/variational

- e.g., GMM + hard clustering gives kmeans
- e.g., some favorite applications:
  - hmm
  - vbmod: arXiv:0709.3512
general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
- e.g., some favorite applications:
  - hmm
  - vbmod: arXiv:0709.3512
  - ebfret: ebfret.github.io
general framework: E+M/variational

- e.g., GMM+hard clustering gives kmeans
- e.g., some favorite applications:
  - hmm
  - vbmod: arXiv:0709.3512
  - ebfret: ebfret.github.io
  - EDHMM: edhmm.github.io
example application: LDA+topics

Figure 27: From Blei 2003
rec engine via CTM

Figure 28: From Blei 2011
recall: recommendation via factoring

\[
\min_{U,V} \sum_{i,j} (r_{ij} - u_i^T v_j)^2 + \lambda_u \|u_i\|^2 + \lambda_v \|v_j\|^2,
\]

Figure 29: From Blei 2011
Maximization of the posterior is equivalent to maximizing the complete log likelihood of $U, V, \theta_{1:j},$ and $R$ given $\lambda_u, \lambda_v$ and $\beta$,

$$L = -\frac{\lambda_u}{2} \sum_i u_i^T u_i - \frac{\lambda_v}{2} \sum_j (v_j - \theta_j)^T (v_j - \theta_j)$$

$$+ \sum_j \sum_n \log \left( \sum_k \theta_{jk} \beta_{k,w_{jn}} \right) - \sum_{i,j} \frac{c_{ij}}{2} (r_{ij} - u_i^T v_j)^2.$$  \hspace{1cm} (7)
CTM: updates for factors

\begin{align*}
  u_i & \leftarrow (VC_i V^T + \lambda_u I_K)^{-1} VC_i R_i & (8) \\
  v_j & \leftarrow (UC_j U^T + \lambda_v I_K)^{-1} (UC_j R_j + \lambda_v \theta_j). & (9)
\end{align*}

Figure 31: From Blei 2011
CTM: (via Jensen’s, again) bound on loss

\[ \mathcal{L}(\theta_j) \geq -\frac{\lambda_n}{2} (v_j - \theta_j)^T (v_j - \theta_j) \]
\[ + \sum_n \sum_k \phi_{jnk} \left( \log \theta_{jk} \beta_{k,w_{jn}} - \log \phi_{jnk} \right) \]
\[ = \mathcal{L}(\theta_j, \phi_j). \quad (10) \]

Figure 32: From Blei 2011
Lecture 5 data product
data science and design thinking

- knowing customer
data science and design thinking

- knowing customer
- right tool for right job
data science and design thinking

- knowing customer
- right tool for right job
- practical matters:
data science and design thinking

- knowing customer
- right tool for right job
- practical matters:
  - munging
data science and design thinking

- knowing customer
- right tool for right job
- practical matters:
  - munging
  - data ops
data science and design thinking

- knowing customer
- right tool for right job
- practical matters:
  - munging
  - data ops
  - ML in prod
Thanks! Thanks MLSS students for your great questions; please contact me @chrishwiggins or chris.wiggins@{nytimes,gmail}.com with any questions, comments, or suggestions!