

Counterfactual Model for Learning

CS6780 – Advanced Machine Learning

Spring 2019

Thorsten Joachims

Cornell University

Reading:

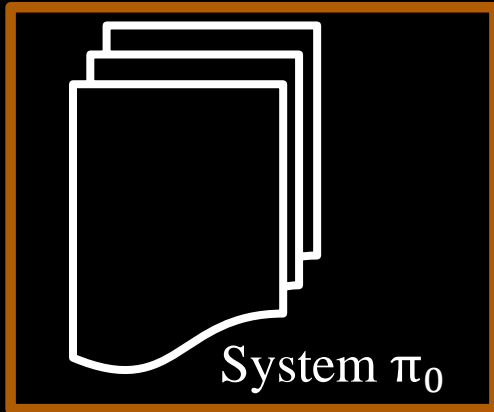
G. Imbens, D. Rubin, Causal Inference for Statistics ..., 2015. Chapters 1,3,12.

Interactive System Schematic



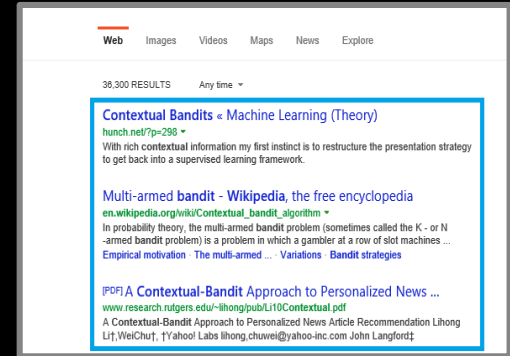
Context x

Feedback $\delta(x, y)$



Utility: $U(\pi_0)$

Action y for x



News Recommender

- Context x :
 - User
- Action y :
 - Portfolio of news articles
- Feedback $\delta(x, y)$:
 - Reading time in minutes



Ad Placement

- Context x :
 - User and page
- Action y :
 - Ad that is placed
- Feedback $\delta(x, y)$:
 - Click / no-click

The screenshot shows a YouTube video player for the video "Frozen Let it Go - In Real Life" by the channel "Working with Lemons". The video is currently playing a scene with Elsa. An advertisement for "MID-YEAR MARVEL DEALS" is displayed on the right side of the page, featuring travel packages to various cities. Below the video, the channel's name, subscriber count (445,097), and video view count (25,728,122) are visible. The page also shows the video's description, publication date (Mar 20, 2015), and a list of comments.

YouTube URL: <https://www.youtube.com/watch?v=hMeiDVv5t8I>

Video Title: Frozen Let it Go - In Real Life

Channel: Working with Lemons (445,097 subscribers)

Views: 25,728,122

Published: Mar 20, 2015

Description: Check out the Official Working with Lemons merchandise at: <http://shop.maker.tv/collections/work...> Thanks to all of our fans, cast and crew and especially Elsa played by Gamry Bagley. Please continue to share, like and subscribe!

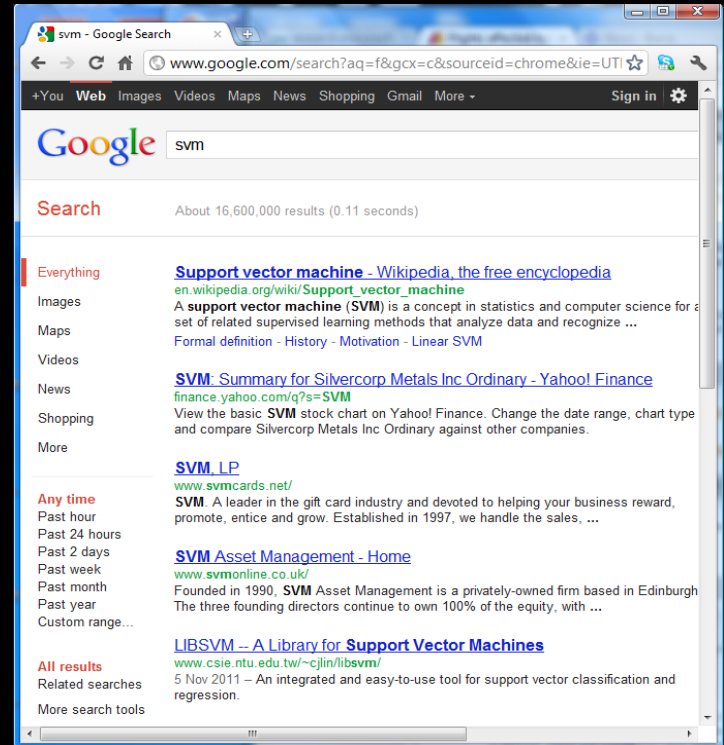
Advertisement: MID-YEAR MARVEL DEALS. FROM HO CHI MINH CITY ECONOMY CLASS. KUALA LUMPUR MELBOURNE AMSTERDAM. 1,731,000 11,248,000 12,978,000. Book: 11 - 29 May 2016. Travel: 14 May - 31 Dec 2016. Terms & Conditions apply. See more deals.

Comments:

- Working with Lemons Shared on Google+ 1 month ago
Let it Go is here!!! Help us share the good news on Facebook and Twitter!
Reply · 103 · 1

Search Engine

- Context x :
 - Query
- Action y :
 - Ranking
- Feedback $\delta(x, y)$:
 - Click / no-click



Log Data from Interactive Systems

- Data

context

π_0 action

reward / loss

$$S = ((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n))$$

→ Partial Information (aka “Contextual Bandit”)
Feedback

- Properties

- Contexts x_i drawn i.i.d. from unknown $P(X)$
- Actions y_i selected by existing system $\pi_0: X \rightarrow Y$
- Feedback δ_i from unknown function $\delta: X \times Y \rightarrow \mathfrak{R}$

Goal

Use interaction log data

$$S = ((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n))$$

- for evaluation of system π

- Offline estimate of online performance of some system π .
- System π can be different from π_0 that generated log.

- for learning new system π

Evaluation: Outline

- Offline Evaluating of Online Metrics
 - A/B Testing (on-policy)
 - Counterfactual estimation from logs (off-policy)
- Approach 1: “Model the world”
 - Imputation via reward prediction
- Approach 2: “Model the bias”
 - Counterfactual model and selection bias
 - Inverse propensity scoring (IPS) estimator

Online Performance Metrics

Example metrics

- CTR
- Revenue
- Time-to-success
- Interleaving
- Etc.

→ Correct choice depends on application and is not the focus of this lecture.

This lecture:

Metric encoded as $\delta(x, y)$ [click/payoff/time for (x,y) pair]

System

- Definition [Deterministic Policy]:
Function

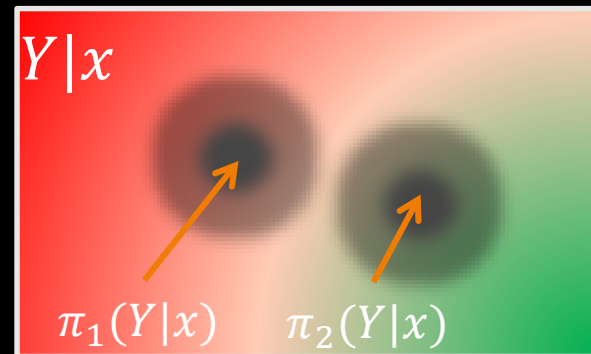
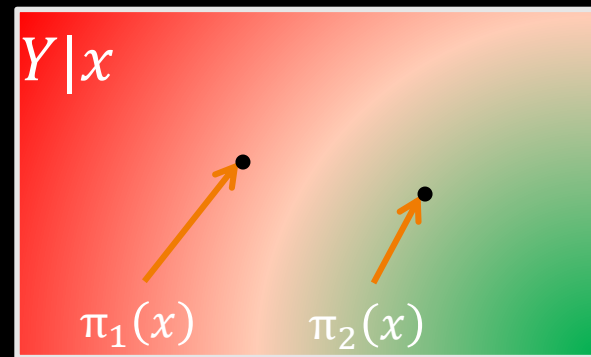
$$y = \pi(x)$$

that picks action y for context x .

- Definition [Stochastic Policy]:
Distribution

$$\pi(y|x)$$

that samples action y given context x

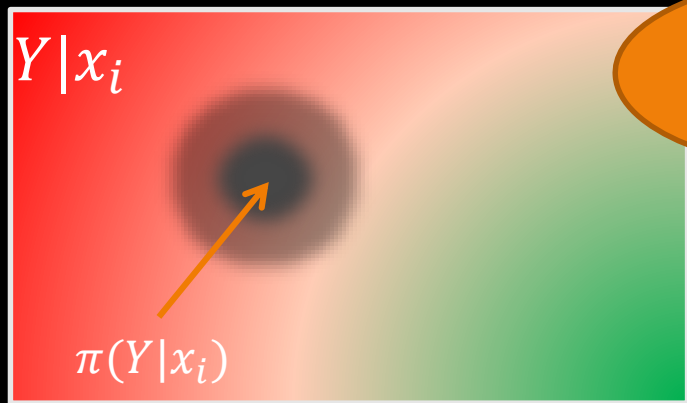


System Performance

Definition [Utility of Policy]:

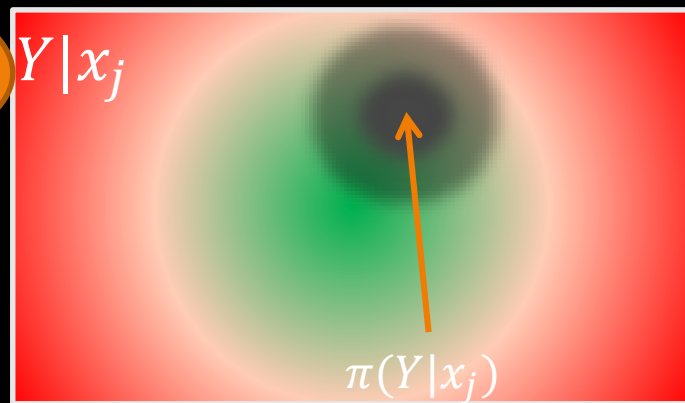
The expected reward / utility $U(\pi)$ of policy π is

$$U(\pi) = \int \int \delta(x, y) \pi(y|x) P(x) dx dy$$



e.g. reading
time of user x
for portfolio y

...



Online Evaluation: A/B Testing

Given $S = ((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n))$ collected under π_0 ,

$$\hat{U}(\pi_0) = \frac{1}{n} \sum_{i=1}^n \delta_i$$

→ A/B Testing

Deploy π_1 : Draw $x \sim P(X)$, predict $y \sim \pi_1(Y|x)$, get $\delta(x, y)$

Deploy π_2 : Draw $x \sim P(X)$, predict $y \sim \pi_2(Y|x)$, get $\delta(x, y)$

⋮

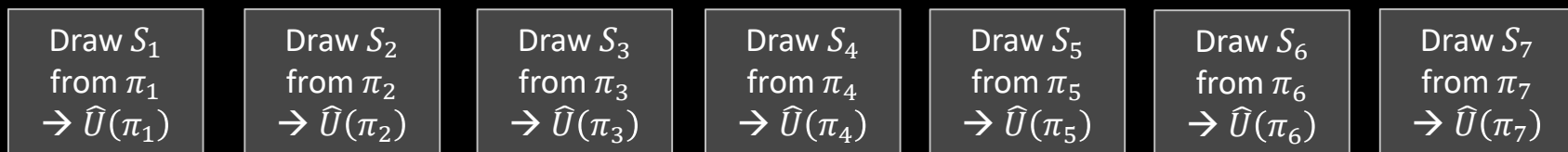
Deploy $\pi_{|H|}$: Draw $x \sim P(X)$, predict $y \sim \pi_{|H|}(Y|x)$, get $\delta(x, y)$

Pros and Cons of A/B Testing

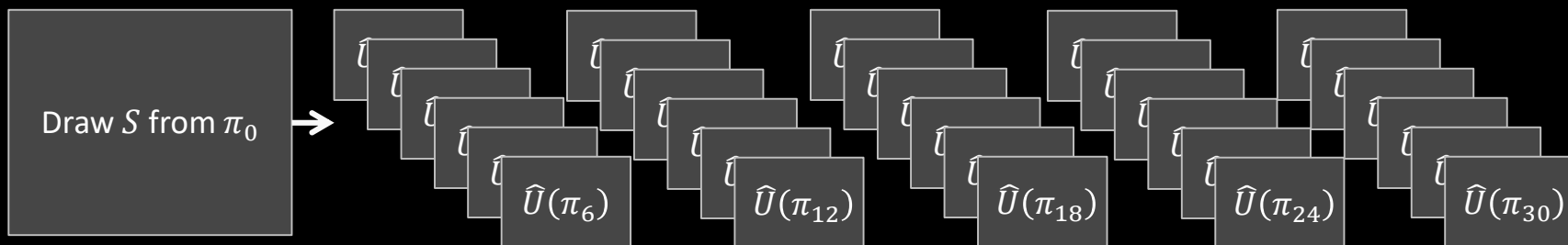
- Pro
 - User centric measure
 - No need for manual ratings
 - No user/expert mismatch
- Cons
 - Requires interactive experimental control
 - Risk of fielding a bad or buggy π_i
 - Number of A/B Tests limited
 - Long turnaround time

Evaluating Online Metrics Offline

- Online: On-policy A/B Test



- Offline: Off-policy Counterfactual Estimates



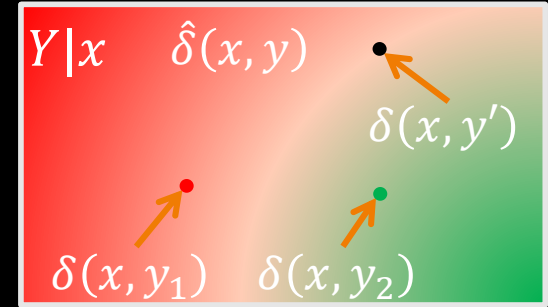
Evaluation: Outline

- Offline Evaluating of Online Metrics
 - A/B Testing (on-policy)
 - Counterfactual estimation from logs (off-policy)
- • Approach 1: “Model the world”
 - Imputation via reward prediction
- Approach 2: “Model the bias”
 - Counterfactual model and selection bias
 - Inverse propensity scoring (IPS) estimator

Approach 1: Reward Predictor

- Idea:

- Use $S = ((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n))$ from π_0 to estimate reward predictor $\hat{\delta}(x, y)$

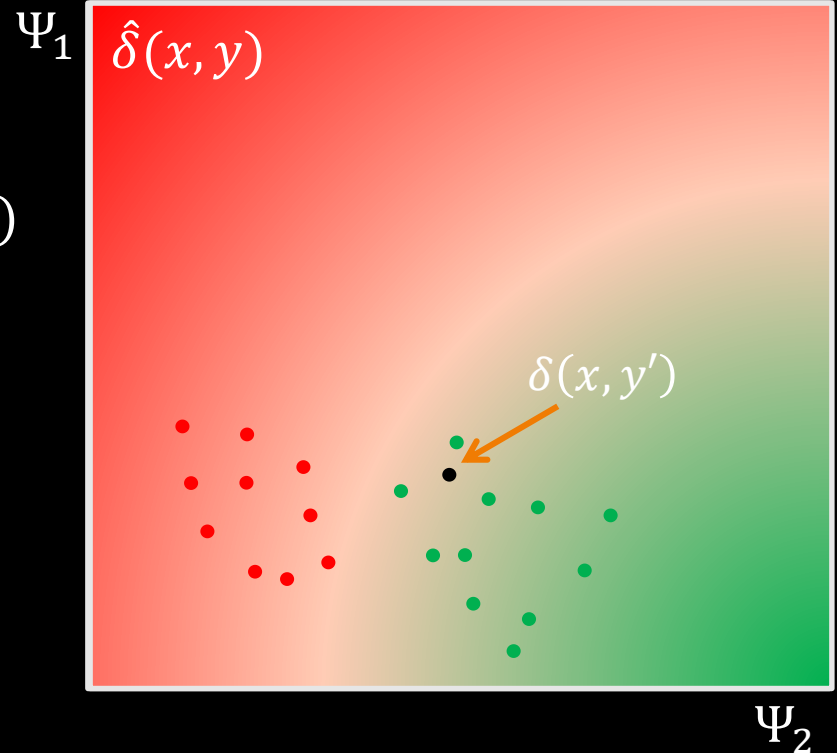


- Deterministic π : Simulated A/B Testing with predicted $\hat{\delta}(x, y)$
 - For actions $y'_i = \pi(x_i)$ from new policy π , generate predicted log $S' = \left((x_1, y'_1, \hat{\delta}(x_1, y'_1)), \dots, (x_n, y'_n, \hat{\delta}(x_n, y'_n)) \right)$
 - Estimate performance of π via $\hat{U}_{rp}(\pi) = \frac{1}{n} \sum_{i=1}^n \hat{\delta}(x_i, y'_i)$
- Stochastic π : $\hat{U}_{rp}(\pi) = \frac{1}{n} \sum_{i=1}^n \sum_y \hat{\delta}(x_i, y) \pi(y|x_i)$

Regression for Reward Prediction

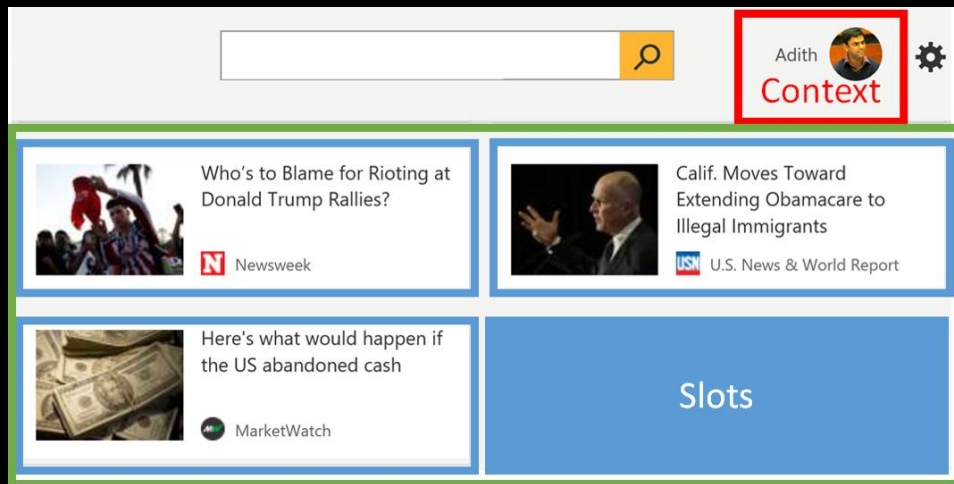
Learn $\hat{\delta}: x \times y \rightarrow \mathfrak{R}$

1. Represent via features $\Psi(x, y)$
2. Learn regression based on $\Psi(x, y)$ from S collected under π_0
3. Predict $\hat{\delta}(x, y')$ for $y' = \pi(x)$ of new policy π

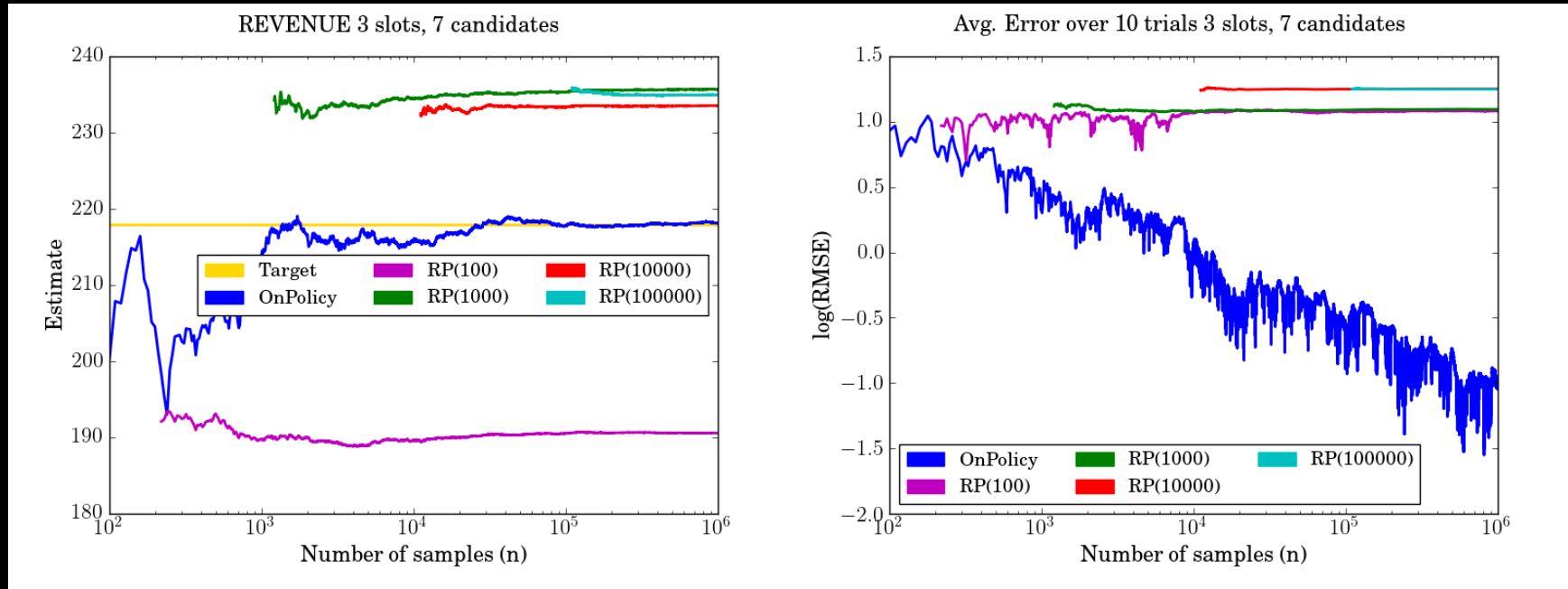


News Recommender: Exp Setup

- Context x : User profile
 - Pick from 7 candidates to place into 3 slots
- Action y : Ranking
 - Complicated hidden function
- Reward δ : “Satisfaction”
 - Placket-Luce “explore around current production ranker”
- Logging policy π_0 : Non-uniform randomized logging system
 - Placket-Luce “explore around current production ranker”



News Recommender: Results



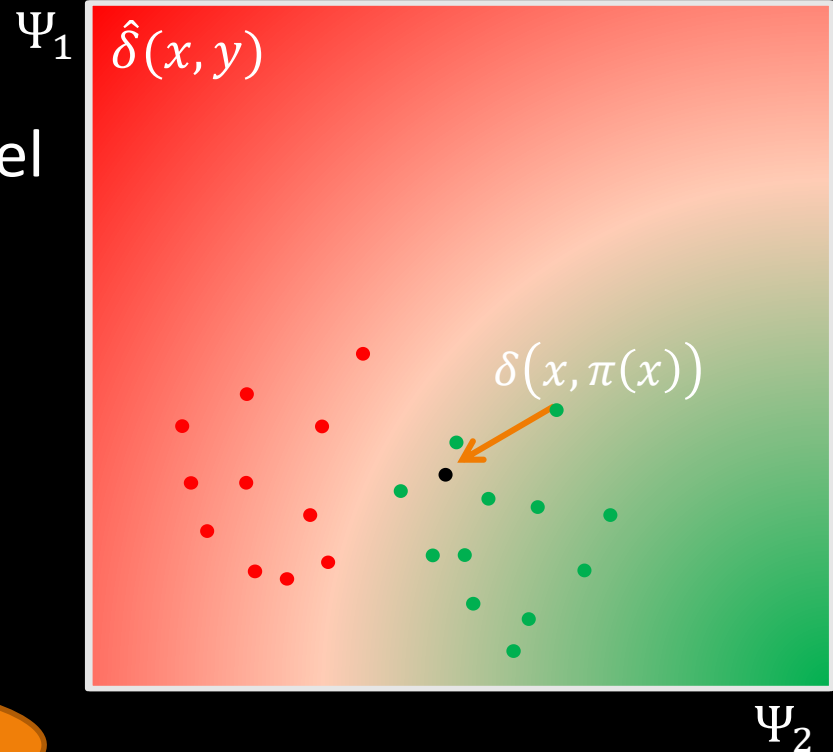
RP is inaccurate even with more training and logged data

Problems of Reward Predictor

- Modeling bias
 - choice of features and model
- Selection bias
 - π_0 's actions are over-represented

$$\rightarrow \hat{U}_{rp}(\pi) = \frac{1}{n} \sum_i \hat{\delta}(x_i, \pi(x_i))$$

Can be unreliable
and biased



Evaluation: Outline

- Offline Evaluating of Online Metrics
 - A/B Testing (on-policy)
 - Counterfactual estimation from logs (off-policy)
- Approach 1: “Model the world”
 - Imputation via reward prediction
- • Approach 2: “Model the bias”
 - Counterfactual model and selection bias
 - Inverse propensity scoring (IPS) estimator

Approach “Model the Bias”

- Idea:

Fix the mismatch between the distribution $\pi_0(Y|x)$ that generated the data and the distribution $\pi(Y|x)$ we aim to evaluate.

$$U(\pi) = \int \int \delta(x, y) \frac{\pi(y|x)}{\pi_0(y|x)} P(x) dx dy$$

Counterfactual Model

- Example: Treating Heart Attacks
 - Treatments: Y
 - Bypass / Stent / Drugs
 - Chosen treatment for patient x_i : y_i
 - Outcomes: δ_i
 - 5-year survival: 0 / 1
 - Which treatment is best?

	Bypass	Stent	Drugs
Patients $x_i \in \{1, \dots, n\}$	0	1	1
		0	1
	1		1
			1
	0	1	0
	1		

Counterfactual Model

Placing Vertical

- Example: ~~Treating Heart Attacks~~
 - Treatments: Y
 - ~~Bypass / Stent / Drugs~~ Pos 1 / Pos 2 / Pos 3
 - Chosen treatment for patient x_i : y_i
 - Outcomes: δ_i
 - ~~5-year survival: 0 / 1~~ Click / no Click on SERP
 - Which treatment is best?

The screenshot shows a Google search for "dogs" with several results. Annotations are placed on the page to identify specific elements for a counterfactual model:

- Pos 1**: Points to the top search result, "Dog - Wikipedia, the free encyclopedia".
- Pos 2**: Points to a news article titled "Pokémon Go has gone to the dogs" from USA TODAY.
- Pos 3**: Points to the bottom search result, "Complete Guide to Caring for Dogs | Dog Breed Information, D." from vetstreet.com.

Handwritten annotations include "S" and "x" near the top of the browser window, and "S" and "x" near the top of the search results area.

Counterfactual Model

- Example: Treating Heart Attacks
 - Treatments: Y
 - Bypass / Stent / Drugs
 - Chosen treatment for patient x_i : y_i
 - Outcomes: δ_i
 - 5-year survival: 0 / 1
 - Which treatment is best?
 - Everybody Drugs
 - Everybody Stent
 - Everybody Bypass
- Drugs 3/4, Stent 2/3, Bypass 2/4 – really?

	Bypass	Stent	Drugs
Patients $x_i, i \in \{1, \dots, n\}$	0	1	1
		0	1
	1		1
			1
		1	0
	1		

Treatment Effects

- Average Treatment Effect of Treatment y

$$- U(y) = \frac{1}{n} \sum_i \delta(x_i, y)$$

- Example

$$- U(\text{bypass}) = \frac{4}{11}$$

$$- U(\text{stent}) = \frac{6}{11}$$

$$- U(\text{drugs}) = \frac{3}{11}$$

	Bypass	Stent	Drugs
0	1	0	
1	1	0	
0	0	1	
0	0	0	
0	1	1	
1	0	0	
1	0	1	
0	1	0	
0	1	0	
1	1	0	
1	1	0	

Assignment Mechanism

- Probabilistic Treatment Assignment

- For patient i : $\pi_0(Y_i = y|x_i)$
- Selection Bias

- Inverse Propensity Score Estimator

- $$\hat{U}_{ips}(y) = \frac{1}{n} \sum_i \frac{\mathbb{I}\{y_i = y\}}{p_i} \delta(x_i, y_i)$$

- Propensity: $p_i = \pi_0(Y_i = y_i|x_i)$

- Unbiased: $E[\hat{U}(y)] = U(y)$,
if $\pi_0(Y_i = y|x_i) > 0$ for all i

- Example

- $$\hat{U}(drugs) = \frac{1}{11} \left(\frac{1}{0.8} + \frac{1}{0.7} + \frac{1}{0.8} + \frac{0}{0.1} \right)$$

$$= 0.36 < 0.75$$

	$\pi_0(Y_i = y x_i)$						
	Bypass	Stent	Drugs		Bypass	Stent	Drugs
Patients	0	1	0	0	1	0	
	1	1	0	1	1	0	
	0	0	1	0	0	1	
	0	0	0	0	0	0	
	0	1	1	0	1	1	
	1	0	0	1	0	0	
	1	0	1	1	0	1	
	0	1	0	0	1	0	
	0	1	0	0	1	0	
	1	1	0	1	1	0	
	1	1	0	1	1	0	

Experimental vs Observational

- Controlled Experiment
 - Assignment Mechanism under our control
 - Propensities $p_i = \pi_0(Y_i = y_i | x_i)$ are known by design
 - Requirement: $\forall y: \pi_0(Y_i = y | x_i) > 0$ (probabilistic)
- Observational Study
 - Assignment Mechanism not under our control
 - Propensities p_i need to be estimated
 - Estimate $\hat{\pi}_0(Y_i | z_i) = \pi_0(Y_i | x_i)$ based on features z_i
 - Requirement: $\hat{\pi}_0(Y_i | z_i) = \hat{\pi}_0(Y_i | \delta_i, z_i)$ (unconfounded)

Conditional Treatment Policies

- Policy (deterministic)
 - Context x_i describing patient
 - Pick treatment y_i based on x_i : $y_i = \pi(x_i)$
 - Example policy:
 - $\pi(A) = \text{drugs}, \pi(B) = \text{stent}, \pi(C) = \text{bypass}$

- Average Treatment Effect

- $$U(\pi) = \frac{1}{n} \sum_i \delta(x_i, \pi(x_i))$$

- IPS Estimator

- $$\hat{U}_{ips}(\pi) = \frac{1}{n} \sum_i \frac{\mathbb{I}\{y_i = \pi(x_i)\}}{p_i} \delta(x_i, y_i)$$

	Bypass	Stent	Drugs	x
	0	1	0	B
	1	1	0	C
	0	0	1	A
	0	0	0	B
	0	1	1	A
	1	0	0	B
	1	0	1	A
	0	1	0	C
	0	1	0	A
	1	1	0	C
	1	1	0	B

Patients

Stochastic Treatment Policies

- Policy (stochastic)
 - Context x_i describing patient
 - Pick treatment y based on x_i : $\pi(Y|x_i)$
- Note
 - Assignment Mechanism is a stochastic policy as well!
- Average Treatment Effect
 - $U(\pi) = \frac{1}{n} \sum_i \sum_y \delta(x_i, y) \pi(y|x_i)$
- IPS Estimator
 - $\hat{U}(\pi) = \frac{1}{n} \sum_i \frac{\pi(y_i|x_i)}{p_i} \delta(x_i, y_i)$

	Bypass	Stent	Drugs	x
	0	1	0	B
	1	1	0	C
	0	0	1	A
	0	0	0	B
	0	1	1	A
	1	0	0	B
	1	0	1	A
	0	1	0	C
	0	1	0	A
	1	1	0	C
	1	1	0	B

Counterfactual Model = Logs

Recorded in Log

Context x_i

Treatment y_i

Outcome δ_i

Propensities p_i

New Policy π

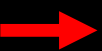
T-effect $U(\pi)$



Average quality of new policy.

Evaluation: Outline

- Evaluating Online Metrics Offline
 - A/B Testing (on-policy)
 - Counterfactual estimation from logs (off-policy)
- Approach 1: “Model the world”
 - Estimation via reward prediction
- Approach 2: “Model the bias”
 - Counterfactual Model
 - Inverse propensity scoring (IPS) estimator



System Evaluation via Inverse Propensity Score Weighting

Definition [IPS Utility Estimator]:

Given $S = ((x_1, y_1, \delta_1), \dots, (x_n, y_n, \delta_n))$ collected under π_0 ,

$$\hat{U}_{ips}(\pi) = \frac{1}{n} \sum_{i=1}^n \delta_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)}$$

Propensity
 p_i

→ Unbiased estimate of utility for any π , if propensity nonzero whenever $\pi(y_i|x_i) > 0$.

Note:

If $\pi = \pi_0$, then online A/B Test with $\hat{U}_{ips}(\pi_0) = \frac{1}{n} \sum_i \delta_i$

→ Off-policy vs. On-policy estimation.

Illustration of IPS

IPS Estimator:

$$\hat{U}_{IPS}(\pi) = \frac{1}{n} \sum_i \frac{\pi(y_i|x_i)}{\pi_0(y_i|x_i)} \delta_i$$

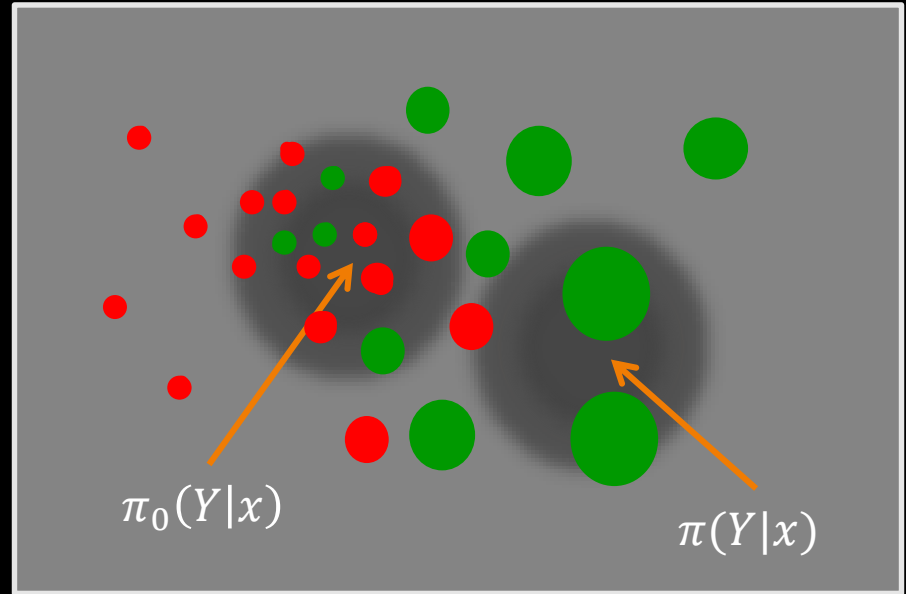
Unbiased:

If

$$\forall x, y: \pi(y|x)P(x) > 0 \rightarrow \pi_0(y|x) > 0$$

then

$$E[\hat{U}_{IPS}(\pi)] = U(\pi)$$



IPS Estimator is Unbiased

$$E[\widehat{U}_{IPS}(\pi)] = \frac{1}{n} \sum_{x_1, y_1} \dots \sum_{x_n, y_n} \left[\sum_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right] \pi_0(y_1 | x_1) \dots \pi_0(y_n | x_n) P(x_1) \dots P(x_n)$$

independent

$$= \frac{1}{n} \sum_{x_1, y_1} \pi_0(y_1 | x_1) P(x_1) \dots \sum_{x_n, y_n} \pi_0(y_n | x_n) P(x_n) \left[\sum_i \frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right]$$

$$= \frac{1}{n} \sum_i \sum_{x_1, y_1} \pi_0(y_1 | x_1) P(x_1) \dots \sum_{x_n, y_n} \pi_0(y_n | x_n) P(x_n) \left[\frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right]$$

marginal

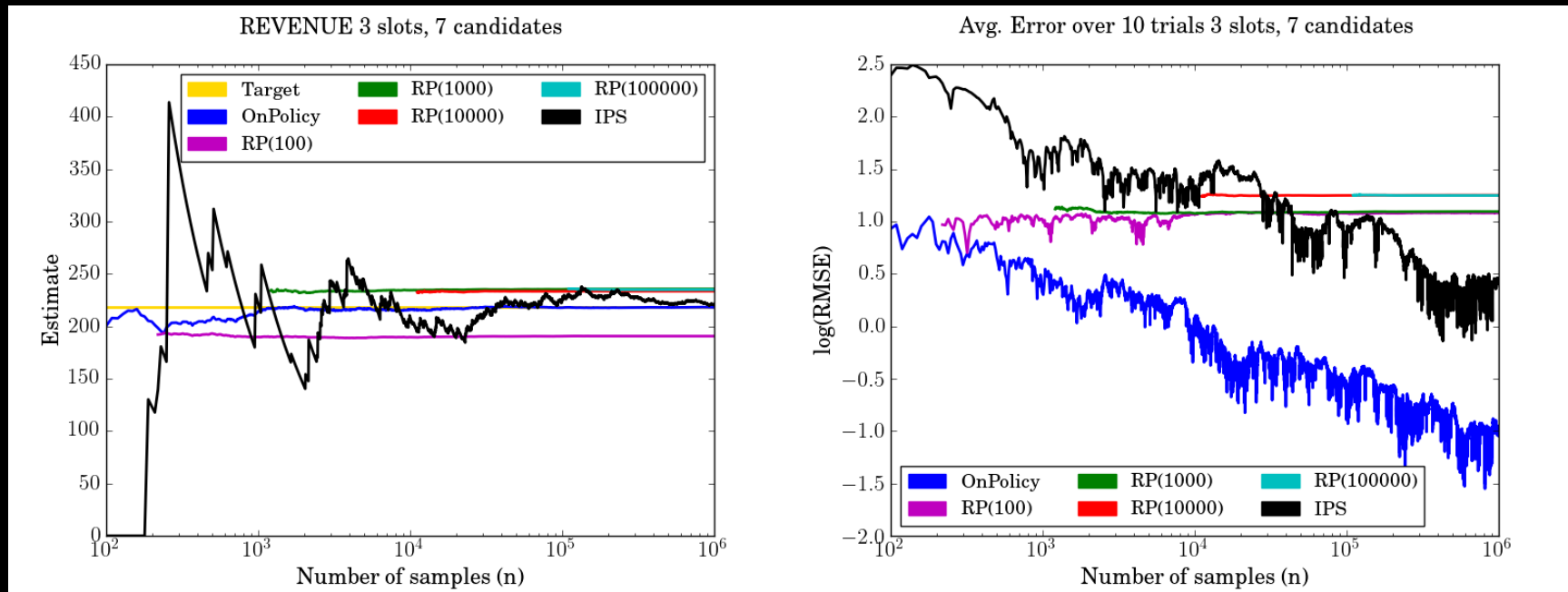
$$= \frac{1}{n} \sum_i \sum_{x_i, y_i} \pi_0(y_i | x_i) P(x_i) \left[\frac{\pi(y_i | x_i)}{\pi_0(y_i | x_i)} \delta(x_i, y_i) \right]$$

full support

$$= \frac{1}{n} \sum_i \sum_{x_i, y_i} P(x_i) \pi(y_i | x_i) \delta(x_i, y_i) = \frac{1}{n} \sum_i U(\pi) = U(\pi)$$

identical x,y

News Recommender: Results



IPS eventually beats RP; variance decays as $O\left(\frac{1}{\sqrt{n}}\right)$

Counterfactual Policy Evaluation

- Controlled Experiment Setting:
 - Log data: $D = ((x_1, y_1, \delta_1, p_1), \dots, (x_n, y_n, \delta_n, p_n))$
 - Observational Setting:
 - Log data: $D = ((x_1, y_1, \delta_1, z_1), \dots, (x_n, y_n, \delta_n, z_n))$
 - Estimate propensities: $p_i = P(y_i | x_i, z_i)$ based on x_i and other confounders z_i
- Goal: Estimate average treatment effect of new policy π .
- IPS Estimator

$$\hat{U}(\pi) = \frac{1}{n} \sum_i \delta_i \frac{\pi(y_i | x_i)}{p_i}$$

or many others.

Evaluation: Summary

- Offline Evaluation of Online Metrics
 - A/B Testing (on-policy)
 - Counterfactual estimation from logs (off-policy)
- Approach 1: “Model the world”
 - Estimation via reward prediction
 - Pro: low variance
 - Con: model mismatch can lead to high bias
- Approach 2: “Model the bias”
 - Counterfactual Model
 - Inverse propensity scoring (IPS) estimator
 - Pro: unbiased for known propensities
 - Con: large variance

From Evaluation to Learning

- Naïve “Model the World” Learning:

- Learn: $\hat{\delta}: x \times y \rightarrow \mathfrak{R}$
- Derive Policy:

$$\pi(y|x) = \operatorname{argmin}_{y'} [\hat{\delta}(x, y')]$$

- Naïve “Model the Bias” Learning:

- Find policy that optimizes IPS training error

$$\pi = \operatorname{argmin}_{\pi'} \left[\sum_i \frac{\pi'(y_i|x_i)}{\pi_0(y_i|x_i)} \delta_i \right]$$