Lecture 12: Batch RL

Emma Brunskill

CS234 Reinforcement Learning.

Winter 2018

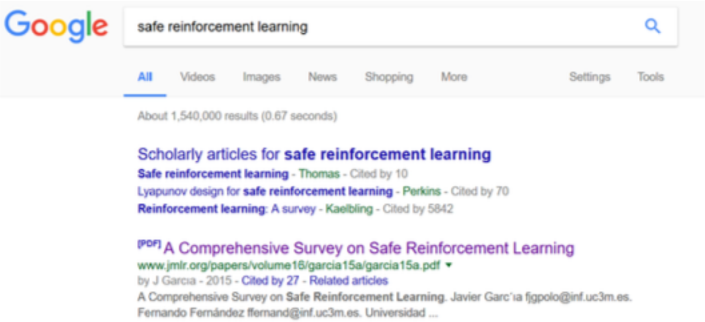
Slides drawn from Philip Thomas with modifications

Class Structure

- Last time: Fast Reinforcement Learning / Exploration and Exploitation
- **This time: Batch RL**
- Next time: Monte Carlo Tree Search

Table of Contents

What does it mean to for a reinforcement learning algorithm to be safe?



Journal of Machine Learning Research 16 (2015) 1437-1480 Submitted 12/13; Revised 11/14; Published 8/15

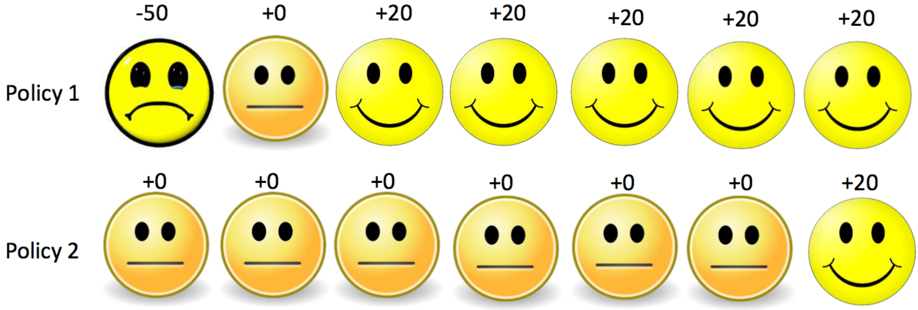
A Comprehensive Survey on Safe Reinforcement Learning

Javier García FJGPOLO@INF.UC3M.ES
Fernando Fernández FFERNAND@INF.UC3M.ES
Universidad Carlos III de Madrid,
Avenida de la Universidad 30,
28911 Leganes, Madrid, Spain

- Safe RL
  - Optimization Criterion
    - Worst Case Criterion
      - Inherent Uncertainty
        - Heger (1994b,a)
        - Gaskett (2003)
      - Parameter Uncertainty
        - Nilim and El Ghaoui (2005)
        - Tamar et al. (2013)
    - Risk-Sensitive Criterion
      - Exponential Functions
        - Howard and Matheson (1972)
        - Borkar (2001, 2002)
        - Basu et al. (2008)
      - Weighted Sum of Return and Risk
        - Mihatsch and Neuneier (2002)
        - Sato et al. (2002)
        - Geibel and Wysotzki (2005)
    - Constrained Criterion
      - Moldovan and Abbeel (2011, 2012a)
      - Castro et al. (2012)
      - Kadota et al. (2006)
    - Other Optimization Criteria
      - Morimura et al. (2010a,b)
      - Luenberger (2013)
      - Castro et al. (2012)
  - Exploration Process
    - External Knowledge
      - Providing Initial Knowledge
        - Driessens and Džeroski (2004)
        - Martín H. and Lope (2009)
        - Song et al. (2012)
      - Deriving a Policy from Demonstrations
        - Abbeel et al. (2010)
        - Tang et al. (2010)
      - Teacher Advice
        - Ask for Help
          - Clouse (1997)
          - García and Fernández (2012)
          - Geramifard et al. (2013)
        - Teacher Provide Advices
          - Clouse and Utgoff (1992)
          - Thomaz and Breazeal (2006, 2008)
          - Vidal et al. (2013)
        - Other Approaches
          - Rosenstein and Barto (2002, 2004)
          - Kuhlmann et al. (2004)
          - Torrey and Taylor (2012)
    - Risk-directed Exploration
      - Gehring and Precup (2013)
      - Law (2005)

Table 1: Overview of the approaches for Safe Reinforcement Learning considered in this survey.

Changing the objective



Changing the objective

- Policy 1:
  - Reward = 0 with probability 0.999999
  - Reward = 10^9 with probability 1-0.999999
  - Expected reward approximately 1000
- Policy 2:
  - Reward = 999 with probability 0.5
  - Reward = 1000 with probability 0.5
  - Expected reward 999.5

Another notion of safety

Safe and efficient off-policy reinforcement learning

Rémi Munos
munos@google.com
Google DeepMind

Thomas Stepleton
stepleton@google.com
Google DeepMind

Anna Harutyunyan
anna.harutyunyan@vub.ac.be
Vrije Universiteit Brussel

Marc G. Bellemare
bellemare@google.com
Google DeepMind

Another notion of safety (Munos et. al)

We start from the recent work of Harutyunyan et al. (2016), who show that naive off-policy policy evaluation, without correcting for the "off-policyness" of a trajectory, still converges to the desired Q^π value function provided the behavior μ and target π policies are not too far apart (the maximum allowed distance depends on the λ parameter). Their $Q^\pi(\lambda)$ algorithm learns from trajectories generated by μ simply by summing discounted off-policy corrected rewards at each time step. Unfortunately, the assumption that μ and π are close is restrictive, as well as difficult to uphold in the control case, where the target policy is greedy with respect to the current Q-function. In that sense this algorithm is not *safe*: it does not handle the case of arbitrary "off-policyness".

Alternatively, the Tree-backup (TB(λ)) algorithm (Precup et al., 2000) tolerates arbitrary target/behavior discrepancies by scaling information (here called *traces*) from future temporal differences by the product of target policy probabilities. TB(λ) is not *efficient* in the "near on-policy" case (similar μ and π), though, as traces may be cut prematurely, blocking learning from full returns.

Another notion of safety

Reachability-Based Safe Learning with Gaussian Processes

Anayo K. Akametalu*
Shahab Kaynama

Jaime F. Fisac*
Melanie N. Zeilinger

Jeremy H. Gillula
Claire J. Tomlin

SAFE REINFORCEMENT LEARNING

A Dissertation Presented

by

PHILIP S. THOMAS

The Problem

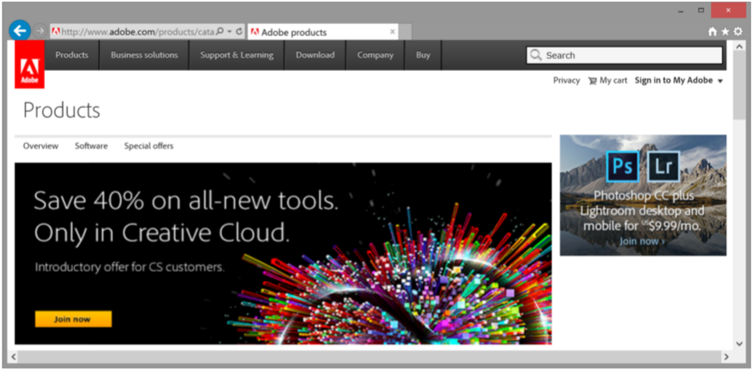
- If you apply an existing method, do you have confidence that it will work?

Reinforcement learning success

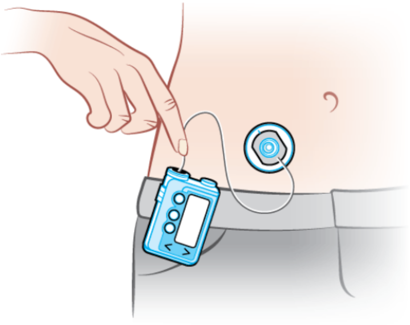
A property of many real applications

- Deploying "bad" policies can be costly or dangerous

Deploying bad policies can be costly



Deploying bad policies can be dangerous



What property should a safe batch reinforcement learning algorithm have?

- Given past experience from current policy/policies, produce a new policy
  - "Guarantee that with probability at least $1 - \delta$, will not change your policy to one that is worse than the current policy."
  - You get to choose δ
  - Guarantee not contingent on the tuning of any hyperparameters

Table of Contents

Notation

- Policy π: $\pi(a) = P(a_t = a \,|\, s_t = s)$
- History: $H = (s_1, a_1, r_1, s_2, a_2, r_2, \cdots, s_L, a_L, r_L)$
- Historical data: $D = \{H_1, H_2, \cdots, H_n\}$
- Historical data from behavior policy, π_b
- Objective:

$$V^{\pi} = \mathbb{E}[\sum_{t=1}^{L} \gamma^t R_t \,|\, \pi]$$

Safe batch reinforement learning algorithm

- Reinforcement learning algorithm, $\mathcal{A}$
- Historical data, D, which is a random variable
- Policy produced by the algorithm, $\mathcal{A}(D)$, which is a random variable
- a safe batch reinforement learning algorithm, $\mathcal{A}$, satisfies:

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V^{\pi_b} \geq 1 - \delta$$

or, in general

$$\Pr(V^{\mathcal{A}(\mathcal{D})} \geq V_{min}) \geq 1 - \delta$$

Table of Contents

Create a safe batch reinforement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1-\delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforement learning algorithm, a

Off-policy policy evaluation (OPE)

Off-policy policy evaluation (OPE)

Importance Sampling (Reminder)

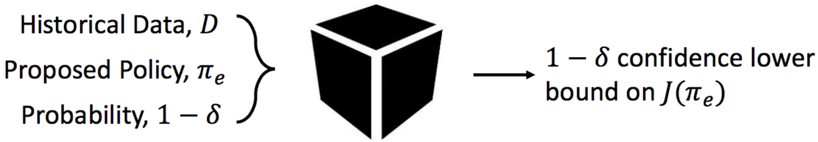
$$IS(D) = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \right) \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$$

$$\mathbb{E}[IS(D)] = V^{\pi_e}$$

Create a safe batch reinforement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1-\delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforement learning algorithm, a

High-confidence off-policy policy evaluation (HCOPE)



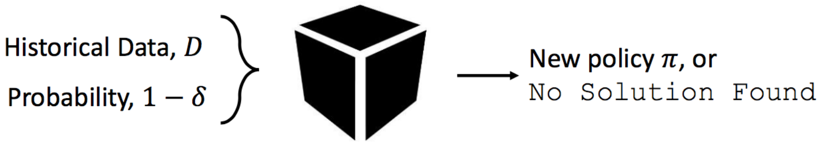
Hoeffding's inequality

- Let $X_1, \cdots, X_n$ be n independent identically distributed random variables such that $X_i \in [0, b]$
- Then with probability at least $1 - \delta$:

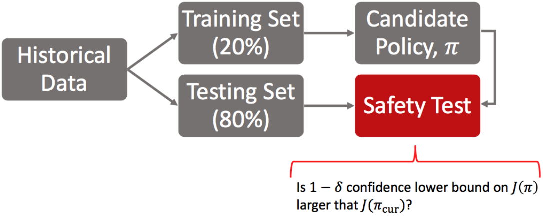
$$\mathbb{E}[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln(1/\delta)}{2n}},$$

where $X_i = \frac{1}{n}\sum_{i=1}^{n}(w_i \sum_{t=1}^{L} \gamma^t R_t^i)$ in our case.

Safe policy improvement (SPI)



Safe policy improvement (SPI)



Create a safe batch reinforement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1 - \delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforement learning algorithm, a

WON'T WORK!

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

$$IS(D) = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \right) \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$$

- Per-decision importance sampling (PDIS)

$$PSID(D) = \sum_{t=1}^{L} \gamma^t \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{\tau=1}^{t} \frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)} \right) R_t^i$$

Off-policy policy evaluation (revisited)

- Importance sampling (IS):

$$IS(D) = \frac{1}{n} \sum_{i=1}^{n} w_i \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$$

- Weighted importance sampling (WIS)

$$WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$$

Off-policy policy evaluation (revisited)

- Weighted importance sampling (WIS)

$$WIS(D) = \frac{1}{\sum_{i=1}^{n} w_i} \sum_{i=1}^{n} w_i \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$$

- NOT unbiased. When $n = 1, \mathbb{E}[WIS] = J(\pi_b)$
- Strongly consistent estimator of V^{π_e}
  - i.e. $\Pr(\lim_{n \to \infty} WIS(D) = V^{\pi_e}) = 1$
  - If
    - Finite horizon
    - One beahvior policy, or bounded rewards

Off-policy policy evaluation (revisited)

- Weighted per-decision importance sampling
  - Also called consistent weighted per-decision importance sampling
  - A fun exercise!

Control variates

- Given: X
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X$
- Unbiased: $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X] = \mu$
- Variance: $Var(\hat{\mu}) = Var(X)$

Control variates

- Given: $X, Y, \mathbb{E}[Y]$
- Estimate: $\mu = \mathbb{E}[X]$
- $\hat{\mu} = X - Y + \mathbb{E}[Y]$
- Unbiased:
  $\mathbb{E}[\hat{\mu}] = \mathbb{E}[X - Y + \mathbb{E}[Y]] = \mathbb{E}[X] - \mathbb{E}[Y] + \mathbb{E}[Y] = \mathbb{E}[X] = \mu$
- Variance:

$$\begin{aligned} Var(\hat{\mu}) &= Var(X - Y + \mathbb{E}[Y]) = Var(X - Y) \\ &= Var(X) + Var(Y) - 2Cov(X, Y) \end{aligned}$$

- Lower variance if $2Cov(X, Y) > Var(Y)$
- We call Y a control variate
- We saw this idea before: baseline term in policy gradient estimation

Off-policy policy evaluation (revisited)

- Idea: add a control variate to importance sampling estimators
  - X is the importance sampling estimator
  - Y is a control variate build from an approximate model of the MDP
    - $\mathbb{E}[Y] = 0$ in this case
  - $PDIS_{CV}(D) = PDIS(D) - CV(D)$
- Called the doubly robust estimator (Jiang and Li, 2015)
  - Robust to (1) poor approximate model, and (2) error in estimates of π_b
    - If the model is poor,the estimates are still unbiased
    - If the sampling policy is unknown, but the model is good, MSE will still be low
  - $DR(D) = PDIS_{CV}(D)$
- Non-recursive and weighted forms, as well as control variate view provided by Thomas and Brunskill (2016)

Off-policy policy evaluation (revisited)

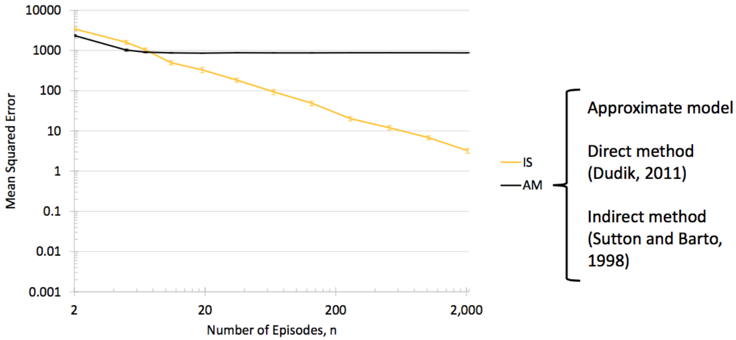
$$DR(\pi_e \mid D) = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t w_t^i (R_t^i - \hat{q}^{\pi_e}(S_t^i, A_t^i)) + \gamma^t \rho_{t-1}^i \hat{v}^{\pi_e}(S_t^i),$$

where $w_t^i = \prod_{\tau 1}^{t} \frac{\pi_e(a_\tau \mid s_\tau)}{\pi_b(a_\tau \mid s_\tau)}$

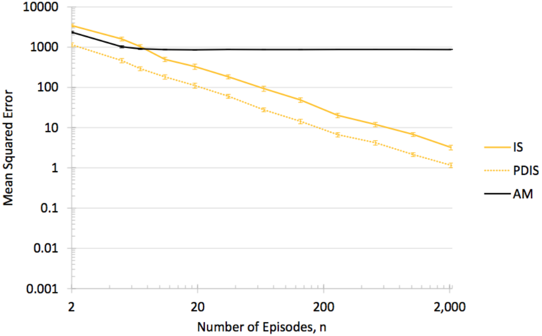
- Recall: we want the control variate Y to cancel with X:

$$R - q(S, A) + \gamma v(S')$$

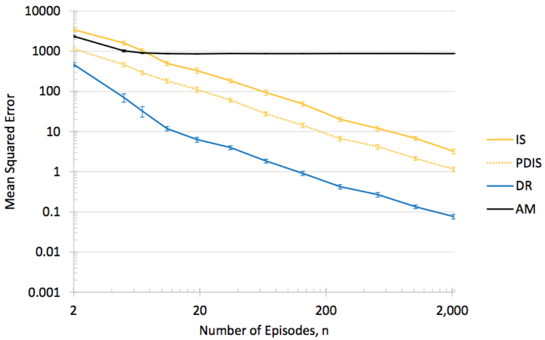
Empirical Results (Gridworld)



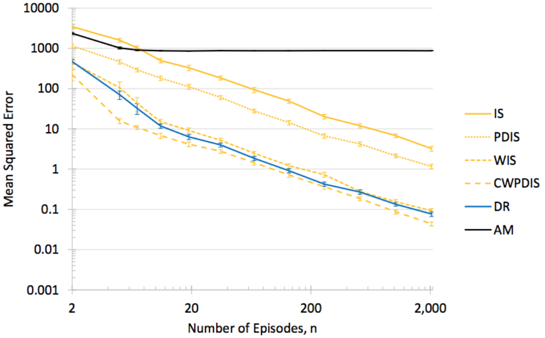
Empirical Results (Gridworld)



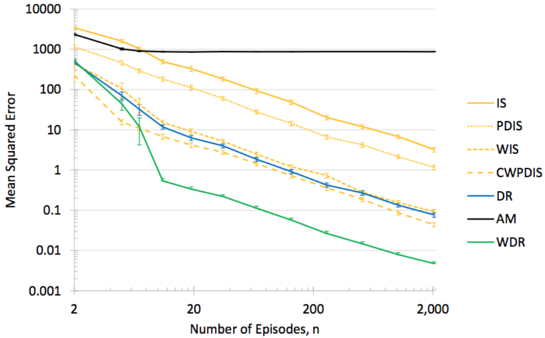
Empirical Results (Gridworld)



Empirical Results (Gridworld)



Empirical Results (Gridworld)

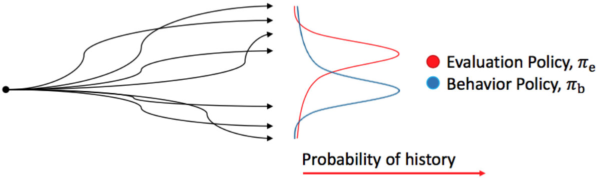


Off-policy policy evaluation (revisited): Blending

- Importance sampling is unbiased but high variance
- Model based estimate is biased but low variance
- Doubly robust is one way to combine the two
- Can also trade between importance sampling and model based estimate within a trajectory
- MAGIC estimator (Thomas and Brunskill 2016)
- Can be particularly useful when part of the world is non-Markovian in the given model, and other parts of the world are Markov

Off-policy policy evaluation (revisited)

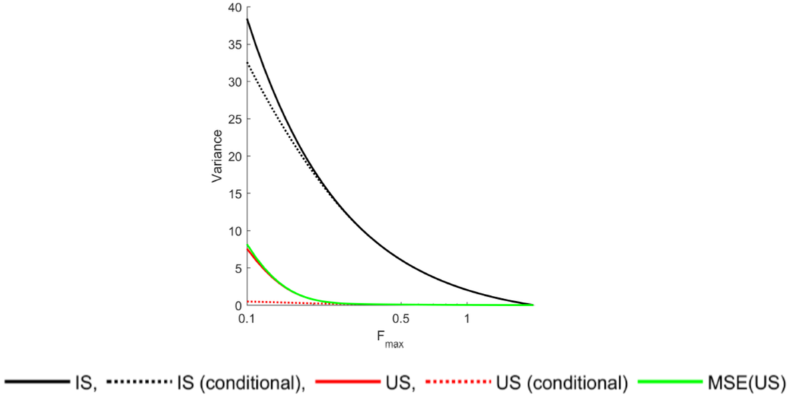
- What if $\text{supp}(\pi_e \subset \text{supp}(\pi_b))$
- There is a state-action pair, (s, a), such that $\pi_e(a \mid s) = 0$, but $\pi_b(a \mid s) \neq 0$.
- If we see a history where (s, a) occurs, what weight should we give it?
- $IS(D) = \frac{1}{n} \sum_{i=1}^{n} \left(\prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)} \right) \left(\sum_{t=1}^{L} \gamma^t R_t^i \right)$



Off-policy policy evaluation (revisited)

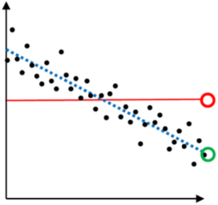
- What if there are zero samples ($n = 0$)?
  - The importance sampling estimate is undefined
- What if no samples are in $\text{supp}(\pi_e)$ (or $\text{supp}(p)$ in general)?
  - Importance sampling says: the estimate is zero
  - Alternate approach: undefined
- Importance sampling estimator is unbiased if $n > 0$
- Alternate approach will be unbiased given that at least one sample is in the support of p
- Alternate approach detailed in Importance Sampling with Unequal Support (Thomas and Brunskill, AAAI 2017)

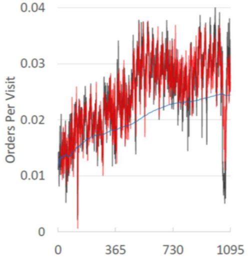
Off-policy policy evaluation (revisited)



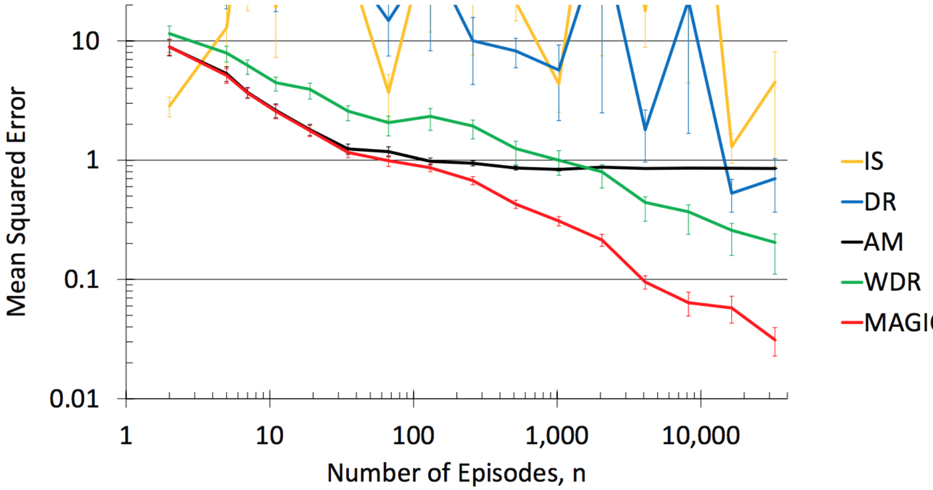
Off-policy policy evaluation (revisited)

- Thomas et. al. Predictive Off-Policy Policy Evaluation for Nonstationary Decision Problems, with Applications to Digital Marketing (AAAI 2017)





Off-policy policy evaluation (revisited)

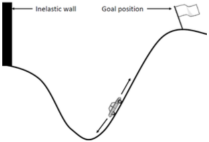


Create a safe batch reinforement learning algorithm

- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1-\delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforement learning algorithm, a

High-confidence off-policy policy evaluation (revisited)

- Consider using IS + Hoeffding's inequality for HCOPE on mountain car



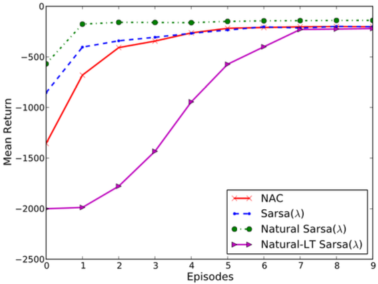
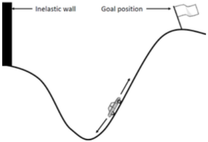


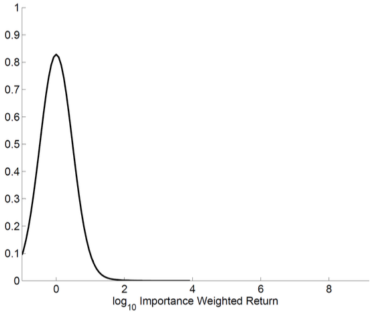
Figure 3: Mountain Car (Sarsa(λ))
Natural Temporal Difference Learning, Dabney and Thomas, 2014

High-confidence off-policy policy evaluation (revisited)

- Using 100,000 trajectories
- Evaluation policy's true performance is $0.19 \in [0, 1]$
- We get a 95% cconfidence lower bound of: -5,8310,000



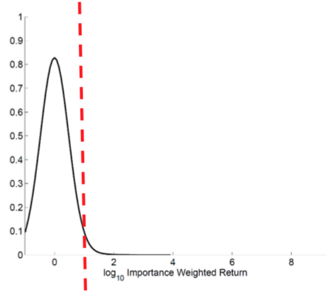
What went wrong



$$w_i = \prod_{t=1}^{L} \frac{\pi_e(a_t \mid s_t)}{\pi_b(a_t \mid s_t)}$$

High-confidence off-policy policy evaluation (revisited)

- Removing the upper tail only decreases the expected value.

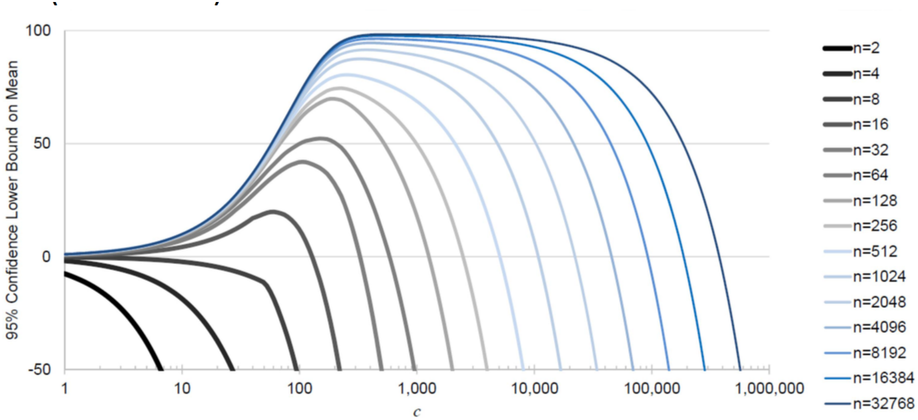


- Thomas et. al, High confidence off-policy evaluation, AAAI 2015

Theorem 1. *Let $X_1, \ldots, X_n$ be n independent real-valued random variables such that for each $i \in \{1, \ldots, n\}$, we have $\mathbb{P}[0 \leq X_i] = 1$, $\mathbb{E}[X_i] \leq \mu$, and some threshold value $c_i > 0$. Let $\delta > 0$ and $Y_i := \min\{X_i, c_i\}$. Then with probability at least $1 - \delta$, we have*

$$\mu \geq \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sum_{i=1}^{n} \frac{Y_i}{c_i}}_{\text{empirical mean}} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \frac{7n \ln(2/\delta)}{3(n-1)}}_{\text{term that goes to zero as } 1/n \text{ as } n \to \infty} - \underbrace{\left(\sum_{i=1}^{n} \frac{1}{c_i}\right)^{-1} \sqrt{\frac{\ln(2/\delta)}{n-1} \sum_{i,j=1}^{n} \left(\frac{Y_i}{c_i} - \frac{Y_j}{c_j}\right)^2}}_{\text{term that goes to zero as } 1/\sqrt{n} \text{ as } n \to \infty}. \tag{3}$$

High-confidence off-policy policy evaluation (revisited)

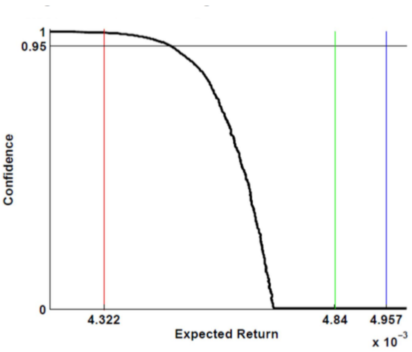


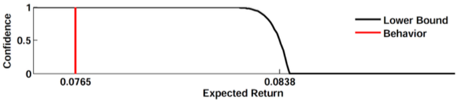
High-confidence off-policy policy evaluation (revisited)

- Use 20% of the data to optimize c
- Use 80% to compute lower bound with optimized c
- Mountain car results:

	CUT	Chernoff-Hoeffding	Maurer	Anderson	Bubeck et al.
95% Confidence lower bound on the mean	0.145	−5,831,000	−129,703	0.055	−.046

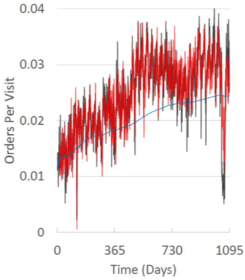
Digital marketing:





High-confidence off-policy policy evaluation (revisited)

Cognitive dissonance:



$$\mathbb{E}[X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i - b\sqrt{\frac{\ln(1/\delta)}{2n}}$$

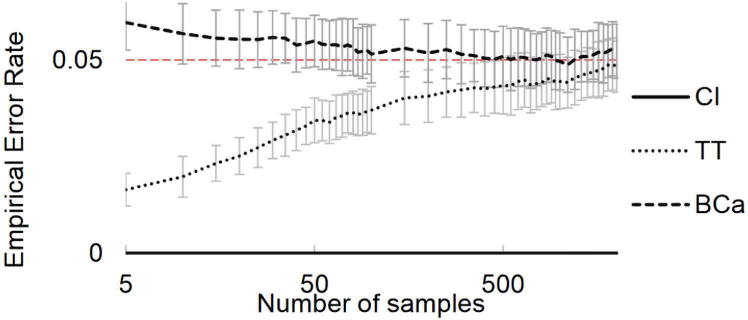
High-confidence off-policy policy evaluation (revisited)

- Student's t-test
  - Assumes that $IS(D)$ is normally distributed
  - By the central limit theorem, it (is as $n \rightarrow \infty$)

$$\Pr\left(\mathbb{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] \geq \frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2}}{\sqrt{n}} t_{1-\delta,n-1}$$
$$\geq 1-\delta$$

  - Efron's Bootstrap methods (e.g., BCa)
    - Also, without importance sampling: Hanna, Stone, and Niekum, AAMAS 2017

High-confidence off-policy policy evaluation (revisited)



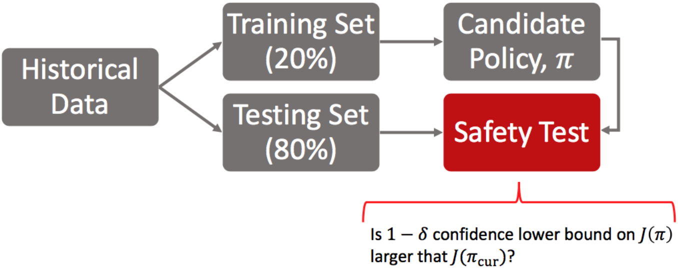
P. S. Thomas. Safe reinforcement learning (PhD Thesis, 2015)

Create a safe batch reinforcement learning algorithm

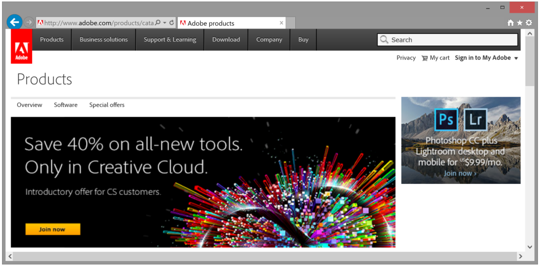
- Off-policy policy evaluation (OPE)
  - For any evaluation policy, π_e, Convert historical data, D, into n independent and unbiased estimates of V^{π_e}
- High-confidence off-policy policy evaluation (HCOPE)
  - Use a concentration inequality to convert the n independent and unbiased estimates of V^{π_e} into a $1-\delta$ confidence lower bound on V^{π_e}
- Safe policy improvement (SPI)
  - Use HCOPE method to create a safe batch reinforcement learning algorithm, a

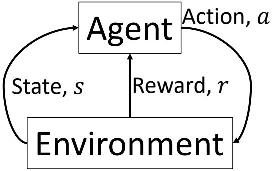
Safe policy improvement (revisited)

Thomas et. al, ICML 2015

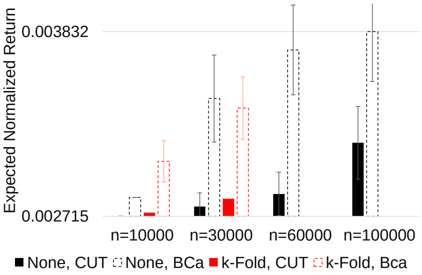


Empirical Results: Digital Marketing

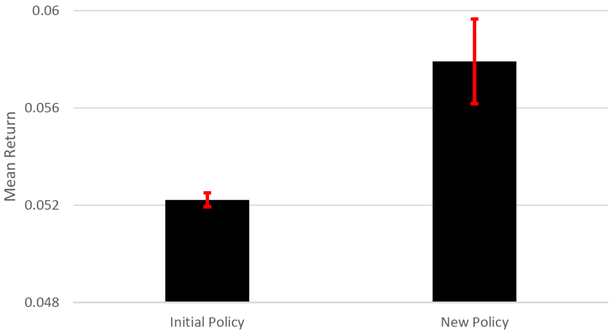




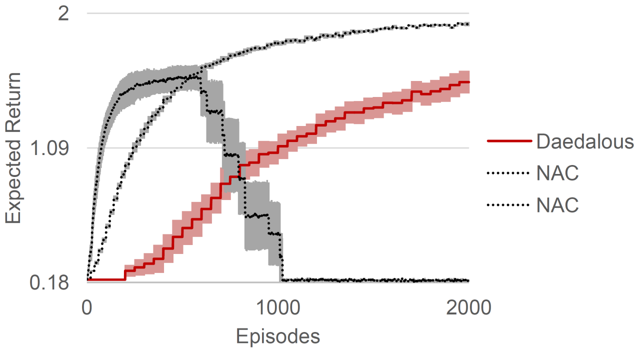
Empirical Results: Digital Marketing



Empirical Results: Digital Marketing

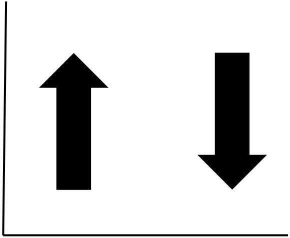


Empirical Results: Digital Marketing

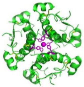


Example Results : Diabetes Treatment

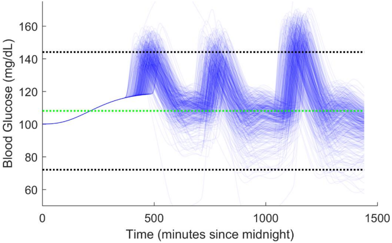
Blood Glucose
(sugar)



Eat Carbohydrates

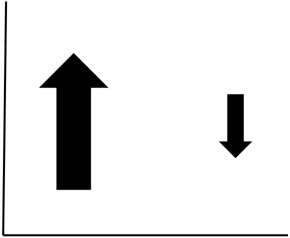


Release Insulin

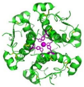


Example Results : Diabetes Treatment

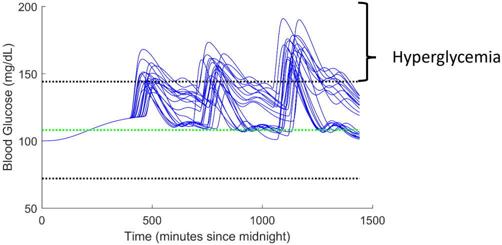
Blood Glucose
(sugar)



Eat Carbohydrates

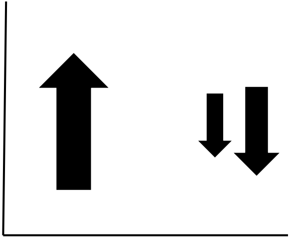


Release Insulin

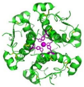


Example Results : Diabetes Treatment

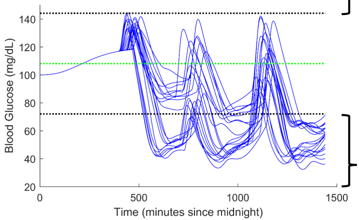
Blood Glucose
(sugar)



Eat Carbohydrates



Release Insulin



Hyperglycemia

Hypoglycemia

Example Results : Diabetes Treatment

$$\text{injection} = \frac{\text{blood glucose} - \text{target blood glucose}}{CF} + \frac{\text{meal size}}{CR}$$

January

S	M	T	W	T	F	S
1	2	3	4	5	6	7
8	9	10	11	12	13	14
15	16	17	18	19	20	21
22	23	24	25	26	27	28
29	30	31				

February

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28				

March

S	M	T	W	T	F	S
			1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	31	

April

S	M	T	W	T	F	S
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30						

May

S	M	T	W	T	F	S
	1	2	3	4	5	6
7	8	9	10	11	12	13
14	15	16	17	18	19	20
21	22	23	24	25	26	27
28	29	30	31			

June

S	M	T	W	T	F	S
				1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	

Example Results : Diabetes Treatment

Alberta Diabetes INSTITUTE

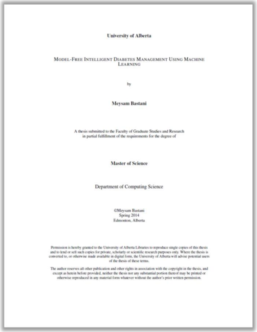
AICML ALBERTA INNOVATES CENTRE FOR MACHINE LEARNING

UNIVERSITY OF ALBERTA

Intelligent Diabetes Management

T1DMS

Type 1 Diabetes Metabolic Simulator



University of Alberta

MODEL-FREE INTELLIGENT DIABETES MANAGEMENT USING MACHINE LEARNING

by

Meysam Bastani

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

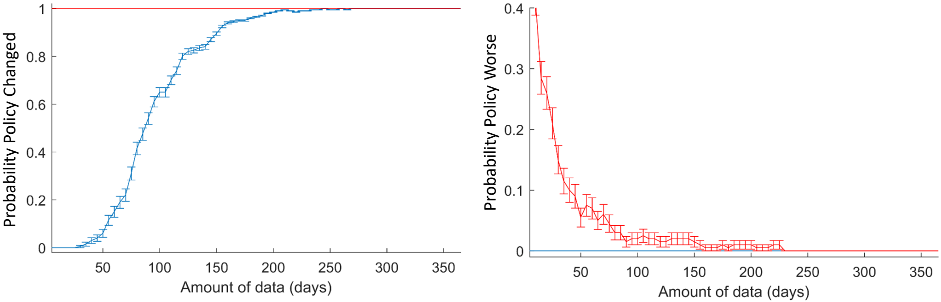
Department of Computing Science

©Meysam Bastani
Spring 2014
Edmonton, Alberta

Permission is hereby granted to the University of Alberta Libraries to reproduce single copies of this thesis and to lend or sell such copies for private, scholarly or scientific research purposes only. Where the thesis is converted to, or otherwise made available in digital form, the University of Alberta will advise potential users of the thesis of these terms.

The author reserves all other publication and other rights in association with the copyright in the thesis, and except as herein before provided, neither the thesis nor any substantial portion thereof may be printed or otherwise reproduced in any material form whatever without the author's prior written permission.

Example Results : Diabetes Treatment



Other Relevant Work

- How to deal with long horizons? (Guo, Thomas, Brunskill NIPS 2017)
- How to deal with importance sampling being "unfair"? (Doroudi, Thomas and Brunskill, best paper UAI 2017)
- What to do when the behavior policy is not known?
- What to do when the behavior policy is deterministic?
- What to do when care about doing safe exploration?
- What to do when care about performance on a single trajectory
- For last two, see great work by Marco Pavone's group, Pieter Abbeel's group, Shie Mannor's group and Claire Tomlin's group, amongst others

Off Policy Policy Evaluation and Selection

- Very important topic: healthcare, education, marketing, ...
- Insights are relevant to on policy learning
- Big focus of my lab
- A number of others on campus also working in this area (e.g. Stefan Wager, Susan Athey...)
- Very interesting area at the intersection of causality and control

What You Should Know: Off Policy Policy Evaluation and Selection

- Be able to define and apply importance sampling for off policy policy evaluation
- Define some limitations of IS (variance)
- List a couple alternatives (weighted IS, doubly robust)
- Define why we might want safe reinforcement learning
- Define the scope of the guarantees implied by safe policy improvement as defined in this lecture

Class Structure

- Last time: Exploration and Exploitation
- **This time: Batch RL**
- Next time: Monte Carlo Tree Search