RL for People

Emma Brunskill
Assistant Professor, Computer Science, Stanford
Thanks to Christoph Dann, Andrea Zanette, Phil Thomas, and Xinkun Nie for some figures & slides
For a doctor does not deliberate whether he shall heal…
assume(s) the end and consider how and by what means it is obtained…
if it is achieved by one means …
consider how it will be achieved… till they come to the first cause…
and what is last in this order of analysis seems to be the first in the order of becoming

Aristotle
Learning to Make Good Sequences of Decisions Under Uncertainty
Learning to Make Good Sequences of Decisions Under Uncertainty ⇒ 1980s Reinforcement Learning
2010s: A New Era of Reinforcement Learning
Mastering the Game of Go Without Human Knowledge
- Silver et al. Nature 2017
Is Reinforcement Learning “Solved”?

Image from DeepMind
Parallel Legacy of “RL” to Benefit People

https://web.stanford.edu/group/cslipublications/cslipublications/SuppesCorpus/Professional%20Photos/albu
- Simulator of domain
- Enormous data to train
- Can always try out a new strategy in domain
- No good simulator of human physiology, behavior & learning
- Gathering real data involves impacting real people

VS

- Simulator of domain
- Enormous data to train
- Can always try out a new strategy in domain
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

And if can learn to make good decisions faster, benefit more people
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Exploration: how to quickly gather information to learn to make good decisions
2. Counterfactual /batch reinforcement learning: leveraging past data
Background: Markov Decision Process

$s_t \in S$
Background: Markov Decision Process

\[ s_t \in S \quad \pi_t(s_t) \rightarrow a_t \]

\[ a_t \in A \]
Background: Markov Decision Process

\[ r(s_t, a_t) \]

\[ s_t \in S \]

\[ \pi_t(s_t) \rightarrow a_t \]

\[ a_t \in A \]
Background: Markov Decision Process Value Function

\[ V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s'|s, a) V^\pi(s') \]

Value func. Reward Dynamics
Background: Markov Decision Process Value Function

$$r(s_t, a_t)$$
$$s_t \in S$$
$$\pi_t(s_t) \rightarrow a_t$$
$$a_t \in A$$

$$V^\pi(s) = r(s, \pi(s)) + \gamma \sum_{s'} p(s' | s, a) V^\pi(s')$$

$$V^*(s) = \max_\pi V^\pi(s)$$
Background: Reinforcement Learning

Only observed through samples (experience)

\[ r(s_t, a_t) \]
\[ s_t \in S \]
\[ \pi_t(s_t) \rightarrow a_t \]
\[ a_t \in A \]

\[
V^\pi(s) = \underbrace{r(s, \pi(s))}_{\text{Value func.}} + \gamma \sum_{s'} p(s'|s, a) V^\pi(s')
\]

Dynamics

Reward

Value function

Only observed through samples (experience)
Background: Probably Approximately Correct RL

\[ r(s_t, a_t) \]
\[ s_t \in S \]
\[ \pi_t(s_t) \rightarrow a_t \]
\[ a_t \in A \]

\[ P \left( \sum_t \mathbb{1} \left( V^{\pi_t}(s_t) < V^*(s_t) - \epsilon \right) \leq F(\epsilon, \delta, S, A) \right) \geq 1 - \delta \]

Sample complexity
Background: Regret Bounds for RL

\[ r(s_t, a_t) \leftarrow a_t \in A \]

\[ s_t \in S \]

\[ \pi_t(s_t) \rightarrow a_t \]

\[ P\left(\sum_t r(s_t, \pi^*(s_t)) - \sum_t r(s_t, \pi_t(s_t)) \leq F(\delta, S, A, T)\right) \geq 1 - \delta \]

Regret
Impact of Strategic Exploration / Efficient RL

- Better decisions on more time steps ⇒ better policies for people
- Enable applications that can’t have many poor decisions to benefit from RL

$$r(s_t, a_t) \quad s_t \in S$$

$$\pi_t(s_t) \rightarrow a_t$$

\[ P\left( \sum_t r(s_t, \pi^*(s_t)) - \sum_t r(s_t, \pi_t(s_t)) \leq F(\delta, S, A, T) \right) \geq 1 - \delta \]

Regret
Strategic Exploration Can Be Essential

- After 50 million actions
- Bottom: Strategic exploration. (Bellemare et al. 2017)
Understanding When it is Hard to Learn to Make Seq. of Decisions Under Uncertainty

- After 50 million actions
- Bottom: Strategic exploration. (Bellemare et al. 2017)
Tabular Episodic RL

Patient 1
Episode 1
Tabular Episodic RL

Patient 1
Episode 1

Patient 2
Episode 2

Patient 3
Episode 3

...
Goal: Policy that Optimizes Expected Reward Over H Steps
MDP Goal: Policy that Optimizes Expected Reward Over $H$ Steps

Bandit Goal: Policy that Optimizes Expected Reward Over 1 Steps

H interactions vs
MDP Goal: Policy that Optimizes Expected Reward Over H Steps

Bandit Goal: Policy that Optimizes Expected Reward Over 1 Steps

Is MDP setting provably harder as H increases? (COLT Open Question, Jiang & Agarwal 2018 conjecture “yes”)
Surprisingly, can give a partial answer of no

Regret scales with $\sqrt{SAK}$ + terms independent of $K/T$

$K = \# \text{ of episodes agent has acted in, each of length } H$
Techniques to Show This Highlight Insights of Recent Progress in Tabular RL
Early Work: Bound Uncertainty Over Dynamics Model Parameters

\[ |Q^*(s, a) - \hat{Q}^*(s, a)| = |p(s, a)^T V^* - \hat{p}(s, a)^T \hat{V}^*| \]

(Assuming no reward error)

\[ Q^*(s, a) = r(s, a) + \gamma \sum_{s'} p(s, a)V^*(s') \]
Early Work: Bound Uncertainty Over Dynamics Model Parameters

\[ |Q^*(s, a) - \hat{Q}^*(s, a)| = |p(s, a)^T V^* - \hat{p}(s, a)^T \hat{V}^*| \approx \frac{H}{\sqrt{n}} \]

(Assuming no reward error)

\[ Q^*(s, a) = r(s, a) + \gamma \sum_{s'} p(s, a) V^*(s') \]
Better: Bound Uncertainty Over Expected Value

\[
| Q^*(s, a) - \hat{Q}^*(s, a) | = | p(s, a)^T V^* - \hat{p}(s, a)^T \hat{V}^* | \leq \frac{H}{\sqrt{n}} \tag{Hoeffding Inequality}
\]

(Assuming no reward error)

\[
\leq \frac{\sigma_{s,a}^{V^*}}{\sqrt{n}} + \frac{H}{n} \tag{Bernstein Inequality}
\]

\[
\sigma_{s,a}^{V^*} = \text{Var}_{s' \sim p(s, a)} V^*
\]

\[
Q^*(s, a) = r(s, a) + \gamma \sum_{s'} p(s, a) V^*(s')
\]
Better: Bound Uncertainty Over Expected Value
And Use to Create New Optimism Bonuses Used for Decision Making

\[ |Q^*(s, a) - \hat{Q}^*(s, a)| = |p(s, a)^T V^* - \hat{p}(s, a)^T \hat{V}^*| \leq \frac{H}{\sqrt{n}} \]
(Assuming no reward error)

(Hoeffding Inequality)

\[ \sigma_{s,a}^{V^*} \leq \frac{\sigma_{s,a}^{V^*}}{\sqrt{n}} + \frac{H}{n} \]
(Bernstein Inequality)

\[ \sigma_{s,a}^{V^*} = \text{Var}_{s' \sim p(s,a)} V^*(s') \]

\[ Q^* = \max_{s,a} \text{Var}_{s' \sim p(s'|s,a)} V^*(s') \]

“Environmental norm” Maillard et al NeurIPS 2014
2: for $k = 1, 2, \ldots$ do
3:     for $t = H, H - 1, \ldots, 1$ do
4:         for $s \in S$ do
5:             for $a \in A$ do
6:                 $\hat{p} = \frac{p_{sum}(s, a)}{n_k(s, a)}$
7:                 $b_{kv}^k = \phi(\hat{p}(s, a), \bar{V}_{t+1}) + \frac{1}{\sqrt{n(s, a)}} \left( \frac{4J + D_p}{\sqrt{n_k(s, a)}} + B_v \| \bar{V}_{t+1} - V_{t+1} \|_2 \hat{p} \right)$
8:                 $Q(a) = \min \{ H - t, \hat{r}_k(s, \hat{\pi}_k(s, t)) + b_{kv}^k(s, a) + \hat{p}^\top \bar{V}_{t+1} + b_{kv}^k \} $
9:         end for
10:     end for
11: end for
2: for $k = 1, 2, \ldots$ do
3:     for $t = H, H - 1, \ldots, 1$ do
4:         for $s \in S$ do
5:             for $a \in A$ do
6:                 $\hat{p} = \frac{p_{sum}(\cdot, s, a)}{n_k(s, a)}$
7:                 $b_{k}^{pv} = \phi(\hat{p}(s, a), \bar{V}_{t+1}) + \frac{1}{\sqrt{n(s, a)}} \left( \frac{4J + D_p}{\sqrt{n_k(s, a)}} + B_v \| \bar{V}_{t+1} - V_{t+1} \|_2 \cdot \hat{p} \right)$
8:             $Q(a) = \min \{H - t, \hat{r}_k(s, \bar{\pi}_k(s, t)) + b_{k}^{r}(s, a) + \hat{p}^\top \bar{V}_{t+1} + b_{k}^{pv} \}$
9:         end for

Empirical estimates of dynamics and reward

Bonus term depends on upper and lower bounds on value function on next time step
2: for \( k = 1, 2, \ldots \) do \\
3: \hspace{1em} for \( t = H, H - 1, \ldots, 1 \) do \\
4: \hspace{2em} for \( s \in S \) do \\
5: \hspace{3em} for \( a \in A \) do \\
6: \hspace{4em} \hat{p} = \frac{p_{sum}(.s, a)}{n_k(s, a)} \\
7: \hspace{5em} b^p_k = \phi(\hat{p}(s, a), \overline{V}_{t+1}) + \frac{1}{\sqrt{n(s, a)}} \left( \frac{4J + B_p}{\sqrt{n_k(s, a)}} + B_v \| \overline{V}_{t+1} - \bar{V}_{t+1} \|_2 \hat{p} \right) \\
8: \hspace{5em} Q(a) = \min\{H - t, \hat{r}_k(s, \hat{\pi}_k(s, t)) + b^r_k(s, a) + \hat{p}^\top \overline{V}_{t+1} + b^p_k \} \\
9: \hspace{4em} \text{end for} \\
10: \hspace{3em} \text{end for} \\
11: \hspace{2em} \text{end for} \\
12: \hspace{1em} \text{end for}

Bonus term depends on upper and lower bounds on value function on next time step

Do Bellman backup w/ bonuses

Empirical estimates of dynamics and reward
2: for $k = 1, 2, \ldots$ do
3:     for $t = H, H - 1, \ldots, 1$ do
4:         for $s \in S$ do
5:             for $a \in A$ do
6:                 $\hat{p} = \frac{p_{\text{sum}}(\cdot, s, a)}{n_k(s, a)}$
7:                 $b_{k}^{pv} = \phi(\hat{p}(s, a), \bar{V}_{t+1}) + \frac{1}{\sqrt{n(s, a)}} \left( \frac{4J + B_p}{\sqrt{n_k(s, a)}} + B_v \| \bar{V}_{t+1} - \bar{V}_{t+1} \|_2, \hat{p} \right)$
8:             $Q(a) = \min\{ H - t, \hat{r}_k(s, \tilde{\pi}_k(s, t)) + b_{k}^r(s, a) + \hat{p}^T \bar{V}_{t+1} + b_{k}^{pv} \}$
9:         end for
10:     $\tilde{\pi}_k(s, t) = \arg \max_a Q(a)$
11:     $\bar{V}_t(s) = Q(\tilde{\pi}_k(s, t))$
12:     $b_{k}^{pv} = \phi(\hat{p}(s, \tilde{\pi}_k(s, t)), \bar{V}_{t+1}) + \frac{1}{\sqrt{n(s, \tilde{\pi}_k(s, t))}} \left( \frac{4J + B_p}{\sqrt{n_k(s, \tilde{\pi}_k(s, t))}} + B_v \| \bar{V}_{t+1} - \bar{V}_{t+1} \|_2, \hat{p} \right)$
13:     $\bar{V}_t(s) = \max\{ 0, \hat{r}_k(s, \tilde{\pi}_k(s, t)) - b_{k}^r(s, \tilde{\pi}_k(s, t)) + \hat{p}^T \bar{V}_{t+1} - b_{k}^{pv} \}$
14: end for
15: end for
16: Evaluate policy $\tilde{\pi}_k$ and update MLE estimates $\hat{p}(\cdot, \cdot)$ and $\hat{r}(\cdot, \cdot)$
17: end for
Main Result

An algorithm with a (high probability) regret bound:

\[ \tilde{O}\left(\min \left[ \sqrt{Q^*SAT}, \sqrt{\frac{G^2}{H}SAT} \right] + S^{1.5}AH^2(H + \sqrt{S}) \right) \]

Problem dependent constants \(G \) & \(Q^*\)
Algorithm is not given \(G\) or \(Q^*\)

Minimax Optimality in dominant terms

\[ Q^* = \max_{s,a} V_{\text{ar}} s' \sim p(s'|s,a) V^*(s') \]

\[ G = \max_{s_0,a_0,\ldots,a_{H-1},s_H} \sum_{i=1}^{H} r_i \]
Our result: matches minimax worst case, doesn’t require mixing (they are in harder infinite H setting)

Our result: doesn’t require us to know span, tighter S dependence
Enhancing Understanding of When it Is Hard to Learn to Act Well

Stochasticity in the Transition Dynamics

Deterministic MDP

\(\tilde{O}(SAH^2)\)

Bandit Like Structure

\(\tilde{O}(\sqrt{SAT} + [\ldots])\)
Enhancing Understanding of When it Is Hard to Learn to Act Well

Stochasticity in the Transition Dynamics

Deterministic MDP

\[ \tilde{O}(SAH^2) \]

Hard Instances Inducing the Lower Bound

\[ p(i, a) = \frac{1}{n} \]
\[ p(\pm i, a) = \frac{1}{2} + c'(a) \]
\[ r(+) = 1 \]
\[ r(-) = 0 \]

\[ \tilde{O}(\sqrt{HSAT} + [\ldots]) \]

Bandit Like Structure

\[ \tilde{O}(\sqrt{SAT} + [\ldots]) \]

Answers part of COLT 2018 open question (Agarwal and Jiang):

No horizon dependence in regret bound for their setting
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Exploration: how to quickly gather information to learn to make good decisions
   a. Minimax bounds for worst case MDP tabular for regret & PAC
   b. Instance dependent bounds for tabular MDP that suggest exploration is easier than expected in many cases
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Exploration: how to quickly gather information to learn to make good decisions
2. Counterfactual / batch RL: reasoning to best leverage existing data
Given ~11k Learners’ Trajectories With Random Action (Levels)

Goal: Learn a New Policy to Maximize Student Persistence
Counterfactual / Batch Off Policy Reinforcement Learning

\[ r(s_t, a_t) \]
\[ s_t \in S \]
\[ \pi_t(s_t) \rightarrow a_t \]
\[ a_t \in A \]

\[ D: \text{Dataset of } n \text{ trajectories } \tau, \tau \sim \pi_b \]
“What If?” Reasoning Given Past Data

Outcome: 92
Outcome: 91
Outcome: 85

?
Data Is Censored

Outcome: 92

Outcome: 91

Outcome: 85
Need for Generalization

Outcome: 92

Outcome: 91

Outcome: 85

?
Growing Interest in Causal Inference & ML

Donald B. Rubin
Educational Testing Service, Princeton, New Jersey

A discussion of matching, randomization, random sampling, and other methods of estimating causal outcomes is presented. The objective is to identify the characteristics of the causal inference. The methods are illustrated with examples from randomized controlled experiments. The results are derived in general, and may be applied to situations in which the treatment is randomized.

Elementary psychological and educational literature has included extensive citations of the use of randomized studies to identify causality. Campbell & Stanley, 1963. The importance of such literature is that it highlights the importance of randomized experiments in identifying causality. Although randomized experiments are not used in non-scientific investigations today, one is led to believe that they are still used in scientific investigations today, and that the results of the experiments are still relevant to the problem at hand. The results of the experiments are still relevant to the problem at hand.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf

Elements of Causal Inference
Foundations and Learning Algorithms
Batch / Counterfactual Policy Optimization:
Pick Policy w/Best Estimated Expected Sum of Rewards

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) \, ds
\]

\(\mathcal{D}\): Dataset of \(n\) trajectories \(\tau\), \(\tau \sim \pi_b\) 
\(\pi\): Policy mapping \(s \rightarrow a\) 
\(S_0\): Set of initial states 
\(\hat{V}^\pi(s, \mathcal{D})\): Estimate \(V(s)\) w/dataset \(\mathcal{D}\)
Challenge: Using Unbiased Value Estimators Can Fail

Doroudi, Thomas, B UAI 2017 best paper

\[ \arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, D) \, ds \]

\( D \): Dataset of \( n \) traj.s \( \tau \), \( \tau \sim \pi_b \)
\( \pi \): Policy mapping \( s \to a \)
\( S_0 \): Set of initial states
\( \hat{V}^\pi(s, D) \): Estimate \( V(s) \) w/dataset \( D \)
Quest: Batch Policy Optimization w/ Generalization Bounds

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) \, ds - \sqrt{\frac{f(\text{VC}((\mathcal{H}_i), \ldots))}{n}}
\]

\(\mathcal{D}\): Dataset of \(n\) trajectories \(\tau, \tau \sim \pi_b\)
\(\pi\): Policy mapping \(s \rightarrow a\)
\(S_0\): Set of initial states
\(\hat{V}^{\pi}(s, \mathcal{D})\): Estimate \(V(s)\) w/dataset \(\mathcal{D}\)
Quest: Batch Policy Optimization w/ Generalization Bounds

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots \}} \int_{s \in S_0} \hat{V}^\pi(s, D) ds - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}
\]

- $D$: Dataset of $n$ trajectories $\tau$, $\tau \sim \pi_b$
- $\pi$: Policy mapping $s \rightarrow a$
- $S_0$: Set of initial states
- $\hat{V}^\pi(s, D)$: Estimate $V(s)$ w/dataset $D$
Quest: Batch Policy Optimization w/ Generalization Bounds

\[ \text{Policy Optimization} \]

\[ \text{Policy Evaluation} \]

\[ \text{Error Bound} \]

\[ \mathcal{D}: \text{Dataset of } n \text{ trajectories } \tau, \tau \sim \pi_b \]
\[ \pi: \text{Policy mapping } s \rightarrow a \]
\[ S_0: \text{Set of initial states} \]
\[ \hat{V}^\pi(s, \mathcal{D}): \text{Estimate } V(s) \text{ w/dataset } \mathcal{D} \]
arg max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, D) ds - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}

Error Bound

\(D\): Dataset of \(n\) trajectories \(\tau\), \(\tau \sim \pi_b\)
\(\pi\): Policy mapping \(s \rightarrow a\)
\(S_0\): Set of initial states
\(\hat{V}^\pi(s, D)\): Estimate \(V(s)\) with dataset \(D\)
Substantial Literature Focuses on 1 Binary Decision: Treatment Effect Estimation from Old Data
Challenge: Covariate Shift
Different Policies $\rightarrow$ Different Actions $\rightarrow$ Different State Distributions

Gottesman et al. Guidelines for reinforcement learning in healthcare. Nature Medicine 2019. Figure by Debbie Maizels/Springer Nature
More Data Efficient Policy Evaluation with Biased Estimators (Thomas and B, ICML 2016)

\[
\text{arg max}_{\pi \in H_i, H_i \in \{H_1, H_2, \ldots\}} \quad \max_{H_i} \quad \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) \, ds - \sqrt{\frac{f(VC(h), \ldots)}{n}}
\]

\( \mathcal{D} \): Dataset of \( n \) trajectories \( \tau \), \( \tau \sim \pi_b \)
\( \pi \): Policy mapping \( s \rightarrow a \)
\( S_0 \): Set of initial states
\( \hat{V}^\pi(s, \mathcal{D}) \): Estimate \( V(s) \) w/dataset \( \mathcal{D} \)
Counterfactual Reasoning for Policy Evaluation

Parametric Models
of dynamics, rewards
or values fit to data

+ Low variance
- Bias (unless realizable)
Counterfactual Reasoning for 1-Step Decision Making

Parametric Models of dynamics, rewards or values fit to data

- Low variance
- Bias (unless realizable)

Importance Sampling correct mismatch of state-action distribution

- Unbiased under certain assumptions
- High variance
Doubly Robust for 1 Step Decision Making

- Smaller variance than importance sampling
- Unbiased if either model realizable or behavior policy known

Brought to multi-armed bandits (Dudik et al. 2011); and RL (Jiang & Li 2016)
Approach: Bias to Save Variance

Parametric Models of dynamics, rewards or values fit to data + Importance Sampling correct mismatch of state-action distribution

Thomas and B 2016
Blend IS-Based & Model Based Estimators to Directly Min Mean Squared Error

- Bias
- Variance

1-step estimate
2-step
H-step
Blend IS-Based & Model Based Estimators to Directly Min Mean Squared Error

Solve quadratic program using estimated bias and variance

\[ w_1 \quad 2-step\quad w_2 \quad \ldots \quad w_H \]
Weighted Doubly Robust and Model & Guided Importance Sampling Estimators

- Under some standard assumptions, proved both estimators consistent

Thomas and B 2016
Gridworld Simulation

Only 10% of the Data is Needed to Learn a Good Estimate New Policy's Value

Thomas and B 2016
Gridworld Simulation: Needed Only 10% of the Data to Learn a Good Estimate of New Policy’s Value

Thomas and B 2016
More Data Efficient Policy Evaluation with Biased Estimators
(Thomas and B, ICML 2016)

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) ds - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}
\]

- $\mathcal{D}$: Dataset of $n$ trajectories $\tau$, $\tau \sim \pi_b$
- $\pi$: Policy mapping $s \rightarrow a$
- $S_0$: Set of initial states
- $\hat{V}^\pi(s, \mathcal{D})$: Estimate $V(s)$ with dataset $\mathcal{D}$
Identifying a Good Policy for Future Use

$$\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^{\pi}(s, \mathcal{D}) ds - \sqrt{\frac{f(\text{VC}(\mathcal{H}_i), \ldots)}{n}}$$

Policy Optimization

Policy Evaluation

Error Bound

\( \mathcal{D} \): Dataset of \( n \) trajectories \( \tau \), \( \tau \sim \pi_b \)

\( \pi \): Policy mapping \( s \rightarrow a \)

\( S_0 \): Set of initial states

\( \hat{V}^{\pi}(s, \mathcal{D}) \): Estimate \( V(s) \) with dataset \( \mathcal{D} \)
Challenge: Good Error Bound Analysis

\[ \arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) \, ds \quad - \quad \sqrt{\frac{f(\text{VC}(\mathcal{H}_i), \ldots)}{n}} \]

\( \mathcal{D} \): Dataset of \( n \) trajectories, \( \tau, \tau \sim \pi_b \)

\( \pi \): Policy mapping \( s \rightarrow a \)

\( S_0 \): Set of initial states

\( \hat{V}^\pi(s, \mathcal{D}) \): Estimate \( V(s) \) with dataset \( \mathcal{D} \)
Challenge: Good Error Bound Analysis

\[
\begin{align*}
\arg \max_{\pi \in \mathcal{H}_i} \quad & \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \\
\text{Policy Optimization} \\
\int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) ds & - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}} \\
\text{Policy Evaluation} \\
\text{Error Bound}
\end{align*}
\]

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large \( n \)

\( \mathcal{D} \): Dataset of \( n \) trajectories \( \tau \), \( \tau \sim \pi_b \)
\( \pi \): Policy mapping \( s \rightarrow a \)
\( S_0 \): Set of initial states
\( \hat{V}^\pi(s, \mathcal{D}) \): Estimate \( V(s) \) w/dataset \( \mathcal{D} \)
Challenge: Good Error Bound Analysis

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large n
- Kernel function & averager approaches (e.g. Ormoneit & Sten 2002) can need # samples exponential in input state dimension

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) ds - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}
\]

- \(\mathcal{D}\): Dataset of \(n\) trajectories \(\tau, \tau \sim \pi_b\)
- \(\pi\): Policy mapping \(s \rightarrow a\)
- \(S_0\): Set of initial states
- \(\hat{V}^\pi(s, \mathcal{D})\): Estimate \(V(s)\) w/dataset \(\mathcal{D}\)
Challenge: Good Error Bound Analysis

- Importance sampling bounds (e.g. Thomas et al, 2015) ignore hypothesis class structure & are typically require very large $n$
- Kernel function & averager approaches (e.g. Ormoneit & Sten 2002) can need # samples exponential in input state dimension
- FQI bounds (e.g. Munos 2003; Munos & Szepesvári 2008; Antos et al., 2008; Lazaric et al., 2012; Farahmand et al., 2009; Maillard et al., 2010; Le, Voloshin, Yue 2019; Chen & Jiang 2019)
  - Require stronger assumptions (realizability and bounds on the inherent Bellman error)
  - If not realizable, these bounds depend on unknown quantities

\[
\arg \max_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) \, ds - \sqrt{\frac{f(\text{VC}(\mathcal{H}_i), \ldots)}{n}}
\]

$\mathcal{D}$: Dataset of $n$ traj.s $\tau$, $\tau \sim \pi_\beta$
$\pi$: Policy mapping $s \rightarrow a$
$S_0$: Set of initial states
$\hat{V}^\pi(s, \mathcal{D})$: Estimate $V(s)$ w/dataset $\mathcal{D}$
Challenge: Good Error Bound Analysis

- Importance sampling bounds (e.g. Thomas et al, 2015)
- Kernel function (e.g. Ormoneit & Sten 2002)
- FQI bounds (e.g. Munos 2003; Munos & Szepesvári 2008; Antos et al., 2008; Lazaric et al., 2012; Farahmand et al., 2009; Maillard et al., 2010; Le, Voloshin, Yue 2019; Chen & Jiang 2019)
  - Require stronger assumptions (realizability and bounds on the inherent Bellman error)
  - If not realizable, FQI bounds depend on unknown quantities
- Primal dual approaches (e.g. Dai, Shaw, Li, Xiao, He, Liu, Chen, Song 2018) are promising and have similar dependencies

\[
\text{arg max}_{\pi \in \mathcal{H}_i} \max_{\mathcal{H}_i \in \{\mathcal{H}_1, \mathcal{H}_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, \mathcal{D}) ds - \sqrt{\frac{f(VC(\mathcal{H}_i), \ldots)}{n}}
\]

- Policy Optimization
- Policy Evaluation
- Error Bound

\( \mathcal{D} \): Dataset of \( n \) trajectories \( \tau \), \( \tau \sim \pi_b \)
\( \pi \): Policy mapping \( s \rightarrow a \)
\( S_0 \): Set of initial states
\( \hat{V}^\pi(s, \mathcal{D}) \): Estimate \( V(s) \) w/dataset \( \mathcal{D} \)
Aim: Strong Generalization Guarantees on Policy Performance, Alternative: Guarantee Find Best in Class Policy

$$\arg \max_{\pi \in \Pi} \int_{s \in S_0} V^\pi(s) ds$$

$\mathcal{D}$: Dataset of $n$ trajectories $\tau$, $\tau \sim \pi_b$
$\pi$: Policy mapping $s \rightarrow a$
$S_0$: Set of initial states
Aim: Strong Generalization Guarantees on Policy Performance, Alternative: Guarantee Find Best in Class Policy

\[
\max_{\pi \in \Pi} \int_{s \in S_0} V^\pi(s) \, ds - \int_{s \in S_0} V^{\hat{\pi} = f(D)}(s) \, ds \leq \sqrt{\frac{f(\kappa(\Pi), \ldots)}{n}}
\]

\[\kappa(\Pi) = \int_{0}^{1} \sqrt{\log N_{dh}(\varepsilon^2, \Pi)} \, d\varepsilon\]

\(D\): Dataset of \(n\) traj.s \(\tau\), \(\tau \sim \pi_b\)
\(\pi\): Policy mapping \(s \rightarrow a\)
\(S_0\): Set of initial states
1st Guarantees on Performance of Policy Choose Vs Best in Class for When to Treat Policies (w/Xinkun Nie & Stefan Wager, arxiv)

\[
\max_{\pi \in \Pi} \int_{s \in S_0} V^\pi(s) \, ds - \int_{s \in S_0} V^{\hat{\pi} = f(D)}(s) \, ds \leq O \left( \left( c_1 \kappa(\Pi) + c_2 + \sqrt{2 \log \left( \frac{1}{\delta} \right)} \right) \sqrt{\frac{V_{\max}}{n}} \right)
\]

\(D\): Dataset of \(n\) trajectories \(\tau\), \(\tau \sim \pi_b\)

\(\pi\): Policy mapping \(s \rightarrow a\)

\(S_0\): Set of initial states

\[
\kappa(\Pi) = \int_0^1 \sqrt{\log N_{dh}(\epsilon^2, \Pi)} \, d\epsilon
\]
Example: Linear Thresholding Policies

Starting HIV treatment as soon as CD4 count dips below 200

Source: https://alv.mizoapp.com/cd4count/
Example: Linear Thresholding Policies

Starting HIV treatment as soon as CD4 count dips below 200

Stopping treatment as soon as health metric above line

Source: https://alv.mizoapp.com/cd4count/
Selecting a When to Treat Policy

\[
\max_{\pi \in \Pi} \int_{s \in S_0} V^\pi(s) ds - \int_{s \in S_0} V^{\hat{\pi} = f(\mathcal{D})}(s) ds \leq O \left( \left( c_1 \kappa(\Pi) + c_2 + \sqrt{2 \log \left( \frac{1}{\delta} \right)} \right) \sqrt{\frac{V_{\max}}{n}} \right)
\]

\( \mathcal{D} \): Dataset of \( n \) trajectories, \( \tau \sim \pi_b \)
\( \pi \): Policy mapping \( s \to a \n\)
\( S_0 \): Set of initial states

\[
\kappa(\Pi) = \int_0^1 \sqrt{\log N_{dh}(\epsilon^2, \Pi)} d\epsilon
\]
Use an Advantage Decomposition

\[ \Delta_\pi := V_\pi - V_0 = \mathbb{E}_0 \left[ \sum_{t=1}^{T} \mathbf{1}_{t \geq \tau_\pi} \left( \mu_{\text{now},t}(S_t) - \mu_{\text{next},t}(S_t) \right) \right] \]

Difference in expected value of following policy \( \pi \) vs. \( \emptyset \)

“Advantage”: Difference in expected value if we start treating now vs. next time-step

[Murphy 2005]
[Kakade 2003]
Use a Doubly Robust Advantage Decomposition

\[ \Delta_\pi := V_\pi - V_0 = \mathbb{E}_0 \left[ \sum_{t=1}^{T} 1_{t \geq T_\pi} (\mu_{\text{now},t}(S_t) - \mu_{\text{next},t}(S_t)) \right] \]

Difference in expected value of following policy \( \pi \) vs. \( \theta \) never acting

“Advantage”: Difference in expected value if we start treating now vs. next time-step

This is a single-step treatment effect estimation problem!

- Estimate treatment effect with a doubly robust estimator given available dataset \( D \)
- Can learn “nuisance” parameters (propensity weights and value function estimates) at a slower rate and still get \( \sqrt{n} \) regret bounds, under various assumptions
Keeping a health metric above 0

*Evolves with brownian motion*

*Treatment nudges it up, but at a cost*

*Always start with treatment ON*

- *Optimal stopping time* of treatment?
- *Unknown* propensity
- Linear Decision Rules
  
  #covariates = 2

- Observe states + noise

Horizon T=10
Keeping a health metric above 0

*Evolves with brownian motion*

*Treatment nudges it up, but at a cost*

*Always start with treatment ON*

- *Optimal stopping time* of treatment?
- *Unknown* propensity
- Linear Decision Rules
  
  
  \#covariates = 2

- Observe states + noise

Horizon T=10
Fitted Q Iteration Policy Less Interpretable
Quest for Batch Policy Optimization with Generalization Guarantees

\[
\underset{\pi \in H_i}{\text{arg max}} \quad \max_{H_i \in \{H_1, H_2, \ldots\}} \quad \int_{s \in S_0} \hat{V}^\pi(s, D) \, ds \quad - \quad \sqrt{\frac{f(VC(h), \ldots)}{n}}
\]

Policy Optimization

Policy Evaluation

Error Bound


Nie, B, Wagner arxiv; Thomas, da Silva, Barto, B, arxiv
Quest for Batch Policy Optimization with Generalization Guarantees

\[
\arg \max_{\pi \in H_i} \max_{H_i \in \{H_1, H_2, \ldots\}} \int_{s \in S_0} \hat{V}^\pi(s, D) \, ds - \sqrt{\frac{f(VC(h), \ldots)}{n}}
\]

Policy Optimization

Policy Evaluation

Error Bound


Nie, B, Wagner arxiv; Thomas, da Silva, Barto, B, arxiv

Many others also purusing: Farahmand, Ghavamzadeh, Jiang, Lazaric, Li, Mannor, Murphy, Munos, Murphy, Pineau, Szepesvari, Yue, ...
→ Much to be done, including to relax common assumptions
Given ~11k Learners’ Trajectories With Random Action (Levels)

Learn a Policy that Increases Student Persistence

(Mandel, Liu, B, Popovic 2014)
Given ~11k Learners’ Trajectories With Random Action (Levels)

Learned a Policy that Increased Student Persistence by +30%

(Mandel, Liu, B, Popovic 2014)
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

1. Understanding how to gather data to quickly learn to make good decisions
   - Matching upper/lower PAC bounds for tabular RL & instance dependent bounds

2. Inferring the outcome of alternate futures
   - More accurate estimators and steps towards strong generalization guarantees
Techniques to Minimize & Understand Data Needed to Learn to Make Good Decisions

⇒ If can learn to make good decisions faster, benefit more people
Better Together? Exploration & Counterfactuals/Off Policy Evaluation (Dann, Wei, Li, B ICML 2019)

- Get tighter confidence bounds!
- Insight: policy deployed is approximation to optimal (static) policy

Output upper/lower bound on policy performance before execute it
Ongoing Work and Open Questions

- Regret analysis using semiparametrics for general Q functions
- Stronger generalization guarantees under weaker assumptions for high dimensional settings
- Nonstationarity
- Standard assumptions -- overlap, no confounding, no interference -- limit these approaches from impacting all the cases want to impact
WDR in Health Example

Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients
Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients

Our weighted DR (WDR) was only consistent off policy estimator tried (PDDDR, PDIS, WPDIS, WDR) that could find an optimal policy which estimated would improve over prior
WDR in Health Example

Sepsis treatment example

(Gottesman et al. arxiv 2018)

- Actions: IV fluids & vasopressors
- Reward: +100 survival, -100 death
- State space: 750 (discretized)
- 19,275 ICU patients

Our weighted DR (WDR) was only consistent off policy estimator tried (PDDDR, PDIS, WPDIS, WDR) that could find an optimal policy which estimated would improve over prior

Under (common) assumption of no confounding, that is not likely to hold in practice