Policy and Off-Policy Evaluation from Data

Mohammad Sadegh Talebi
m.shahi@di.ku.dk
Department of Computer Science

Motivation

We've studied planning in a known discounted MDP:

- Using `VI`, `PI`, and their variants
- *Planning* is a slang word for 'solving MDP'

What if the MDP is unknown but accessible only through collected data?

- RL deals with (near-)optimally solving an unknown MDP using offline/online data (experience).
- The first step is policy evaluation using offline/online data.

PE vs. OPE vs. OPO

Policy Evaluation (PE) from data: Estimate V^π using data sampled from π.

Two related problems:

- **Off-Policy Evaluation (OPE):** Estimate V^π using data collected according to some fixed policy $\pi_\mathrm{b} \neq \pi$
  - π_b is called the behavior (or logging policy) — an exploratory policy.
  - $\pi \neq \pi_\mathrm{b}$ is called the target policy (a.k.a. estimation policy).

- **Off-Policy Optimization (OPO):** Find an optimal policy using data collected according to some behavior policy π_b

OPE/OPO

Consider a company selling products according to some policy A.

- Interactions with the world can be modeled as an MDP.
- The transition function (determined by, e.g., customer arrivals, market dynamics) is unknown, but the company has a rich dataset logged via A.
- The expected revenue under A can be found by computing V^A (Policy Evaluation, *this lecture!*).

Shall the company switch to a new policy B or not?

- Yes, if B yields a higher revenue, i.e., $V^B > V^A$
- One can find the unknown V^B via the dataset of A (via OPE methods).
- Also OPE gives confidence sets on $V^B \implies$ Better to switch to B only if

$$V^B \geq V^A + \text{margin}, \quad \text{with high probability}$$

Part 1: Policy Evaluation

Policy Evaluation

Policy Evaluation

Given: A dataset $\mathcal{D}$ collected under some *fixed* policy π.

Mathematically, $\mathcal{D} = \Big\{(s_t, a_t, r_t), 1 \le t \le n\Big\}$ where

$$a_t \sim \pi(\cdot|s_t), \quad r_t \sim R(s_t, a_t), \quad s_{t+1} \sim P(\cdot|s_t, a_t)$$

Goal: Derive (point) estimate, and possibly confidence intervals, for V^π.

We study two algorithms:

- A model-based method, which we call `MB-PE`.
- A model-free method called Temporal Difference (`TD`) Learning.

`MB-PE`: A Model-Based Method

Known Model

Recall the definition of V^π, $\pi \in \Pi^{\mathsf{SR}}$:

$$V^\pi(s) = \mathbb{E}^\pi\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right]$$

and the Bellman equation:

$$V^\pi(s) = r_t + \gamma \mathbb{E}^\pi\left[\sum_{t=2}^{\infty} \gamma^{t-1} r_t \middle| s_1 = s\right] = r_t + \gamma \mathbb{E}^\pi\left[V^\pi(s_{t+1}) \middle| s_1 = s\right]$$

π induces an MRP (P^π, r^π) with:

$$P^\pi_{s,s'} = \sum_a \pi(a|s) P(s'|s,a), \qquad r^\pi(s) = \sum_a \pi(a|s) r(s,a)$$

$$\text{Then,} \quad V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$$

MB-PE: Idea

Idea: Define empirical estimates for P^π and apply the certainty equivalence principle.

Smoothed Estimator for P^π:

$$\widehat{P}^\pi_{s,s'} = \frac{N(s,s') + \alpha}{N(s) + \alpha S}, \quad \text{with}$$

$$N(s,s') = \sum_{t=1}^{n-1} \mathbb{I}\{s_t = s, s_{t+1} = s'\} \quad \text{and} \quad N(s) = \sum_{s' \in \mathcal{S}} N(s,s')$$

- $\alpha \geq 0$ is an arbitrary choice controlling the level of smoothing.
- $\alpha = 0$ corresponds to Maximum Likelihood Estimator (unbiased).
- $\alpha = 1/S$ corresponds to Laplace Smoothed Estimator (biased, but the bias vanishes as $N(s)$ increases).
- Consistency: $\widehat{P}^\pi_{s,s'}$ converges to $P_{s,s'}$ as $N(s) \to \infty$ almost surely.

MB-PE: Idea

Idea: Define empirical estimates for P^π and apply the certainty equivalence principle.

Smoothed Estimator for r^π:

$$\widehat{r}^\pi(s) = \frac{\alpha + \sum_{t=1}^{n-1} r_t \mathbb{I}\{s_t = s\}}{\alpha + N(s)}$$

- Consistency: $\widehat{r}^\pi(s)$ converges to $r^\pi(s)$ as $N(s) \to \infty$ almost surely.
- Unbiased for $\alpha = 0$.

Then, the following is an estimate for V^π:

$$\widehat{V}^\pi = (I - \gamma \widehat{P}^\pi)^{-1} \widehat{r}^\pi$$

MB-PE: Convergence

Theorem

If all states are visited infinitely often under π, then $\widehat{V}^\pi$ converges to V^π almost surely:

$$\mathbb{P}\Big(\lim_{n\to\infty} \widehat{V}^\pi = V^\pi\Big) = 1$$

- In other words, if π is exploratory enough, $\widehat{V}^\pi$ converges to V^π in the following sense:

$$\mathbb{P}\left(\exists \mathcal{D}, \exists s \in \mathcal{S} : \lim_{t\to\infty} \widehat{V}^\pi(s;\mathcal{D}) \neq V^\pi(s)\right) = 0$$

  I.e., datasets for which $\widehat{V}^\pi \neq V^\pi$ will occur with probability 0.
- It follows from the a.s. convergence of $\widehat{P}^\pi$ to P^π and of $\widehat{r}^\pi$ and r^π.
- We can use concentration inequalities (e.g., Hoeffding's) to derive confidence interval(s) for V^π.

MB-PE: Pros and Cons

This is a model-based approach since it maintains an approximate model of MDP (or MRP) and then computes V^π for that.

Disadvantages of the model-based solution:

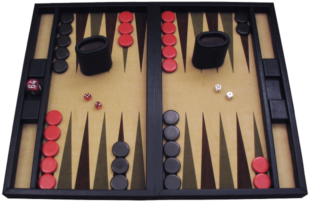
- It results in value estimates with a large variance in practice, which is undesirable.
- It maintains estimates of $S^2 + S$ elements of MRP, though we need to maintain S estimates to find V^π.
- Computational complexity is $O(S^3)$, and space complexity is $O(S^2)$.
- May not be easily converted into an incremental procedure.

Temporal Difference Learning

Temporal Difference Learning

- Temporal Difference Learning was popularized and extended by Richard Sutton in 1988.
- However, the earliest reported use dates back to Arthur Samuel (1959).

Application to Backgammon game by Gerald Tesauro (TD-Gammon), read more here.



source: Wikipedia

Temporal Difference Learning

Assume $\widehat{V}$ is some estimate for V^π — Hence, $\widehat{V}(s_t)$ is an estimate for $V^\pi(s_t)$.

Now consider $r_t + \gamma\widehat{V}(s_{t+1})$:

$$\mathbb{E}\left[r_t + \gamma\widehat{V}(s_{t+1})\middle|s_t, \widehat{V}\right] = \mathbb{E}_{a\sim\pi(s_t)}\left[R(s_t, a) + \gamma\sum_{s'} P(s'|s_t, a)\widehat{V}(s')\middle|s_t, \widehat{V}\right]$$

Hence, $r_t + \gamma\hat{V}(s_{t+1})$ gives *another estimate* for $V^\pi(s_t)$.

Temporal Difference Learning

Ideally we would like to have an estimate $\widehat{V}$ so that:

$$\widehat{V}(s_t) \approx r_t + \gamma \widehat{V}(s_{t+1})$$

- Given $\widehat{V}(s_t)$, in view of Bellman's equation $r_t + \gamma \widehat{V}(s_{t+1})$ serves as a target estimate for $V^\pi(s_t)$.
- The temporal difference error is $\delta_t = r_t + \gamma \widehat{V}(s_{t+1}) - \widehat{V}(s_t)$.

Hence, we may update $\widehat{V}(s_t)$ to reduce the error δ_t:

$$\underbrace{\widehat{V}(s_t)}_{\text{new value}} \longleftarrow \underbrace{\widehat{V}(s_t)}_{\text{old value}} + \alpha_t \underbrace{\Big(r_t + \gamma \widehat{V}(s_{t+1}) - \widehat{V}(s_t)\Big)}_{\text{estimation error}}$$

This method is called Temporal Difference (TD) learning — this is a form of bootstrapping, since we refined $\widehat{V}(s_t)$ using another estimate.

TD: Learning Rate

To guarantee convergence, learning rates $(\alpha_t)_{t\geq 1}$ must satisfy the *Robbins-Monro conditions*:

$$\alpha_t > 0, \qquad \sum_{t=1}^{\infty} \alpha_t = \infty, \qquad \sum_{t=1}^{\infty} \alpha_t^2 < \infty$$

(I.e., a positive sequence that is *square-summable-but-not-summable.*)

Examples:

- $\alpha_t = \frac{1}{t+1}$
- $\alpha_t = \frac{2}{\sqrt{t}\log(t+1)}$
- $\alpha_t = \frac{c}{t^a}$ for $a \in (\frac{1}{2}, 1]$ and $c > 0$

TD

- **input:** $\mathcal{D} = \{(s_t, r_t)\}_{1\le t\le n}$, $(\alpha_t)_{t\ge 1}$
- **initialization:** Select V_1 arbitrarily
- **for** $t = 1, \dots, n-1$ Update:

$$V_{t+1}(s) = \begin{cases} V_t(s) + \alpha_t\Big(r_t + \gamma V_t(s_{t+1}) - V_t(s)\Big) & s = s_t \\ V_t(s) & \text{else.} \end{cases}$$

- **output:** V_n

TD: Advantages

- TD is model-free: It does not require a model of the MDP, only relies on collected experience.
- TD can be incremental (unlike the model-based methods).
- Computational complexity (per-step) is $O(1)$. Space complexity is S. Much cheaper than the model-based method.
- TD results in estimates for V^π with low variance.

Is TD Gradient?

- TD update resembles Stochastic Gradient Descent (SGD).
- However, it can be shown that TD is not an SGD for *any objective function* (see Philip Thomas' Notes, p. 69).
- In fact, TD is a Stochastic Approximation (SA) algorithm and it inherits convergence guarantee from SA — we briefly overview SA in next lecture.

TD: Convergence

Theorem

If all states are visited infinitely often under π and $(\alpha_t)_{t\geq 1}$ satisfies the Robbins-Monro conditions, then V_t converges to the true value function V^π almost surely:

$$\mathbb{P}\left(\forall s \in \mathcal{S},\ \lim_{t\to\infty} V_t(s) = V^\pi(s)\right) = 1$$

In other words, if π is exploratory enough, V_t converges to V^π, in the following sense:

$$\mathbb{P}\left(\exists \mathcal{D}, \exists s \in \mathcal{S} : \lim_{t\to\infty} V_t(s;\mathcal{D}) \neq V^\pi(s)\right) = 0$$

I.e., datasets for which $V_\infty \neq V^\pi$ will occur with probability 0.

TD(λ)

TD only uses only r_t and $\widehat{V}(s_{t+1})$ to refine $\widehat{V}(s_t)$ — i.e., it looks *one-step into future*.

Why not looking into ℓ-step into future? using the target

$$\sum_{n=0}^{\ell} \gamma^n r_{t+n} + \gamma^{\ell+1} \widehat{V}(s_{t+\ell+1})$$

The temporal difference error when using ℓ-step lookahead is:

$$\begin{aligned}\delta_t^\ell &= \sum_{n=0}^{\ell} \gamma^n r_{t+n} + \gamma^{\ell+1} \widehat{V}(s_{t+\ell+1}) - \widehat{V}(s_t) \\ &= \sum_{n=0}^{\ell} \gamma^n \Big(r_{t+n} + \gamma \widehat{V}(s_{t+n+1}) - \widehat{V}(s_{t+n}) \Big)\end{aligned}$$

TD(λ)

Looking into ℓ-step into future:

Now let's update $\widehat{V}(s_t)$ using a mixture of ℓ-steps information each weighted with $(1-\lambda)\lambda^\ell$ for some $\lambda \in [0,1)$:

$$\begin{aligned}\widehat{V}(s_t) &\longleftarrow \widehat{V}(s_t) + \alpha_t \sum_{\ell=0}^{\infty}(1-\lambda)\lambda^\ell \delta_t^\ell \\ &= \widehat{V}(s_t) + \alpha_t \sum_{n=0}^{\infty} \lambda^n \gamma^n \Big(r_{t+n} + \gamma \widehat{V}(s_{t+n+1}) - \widehat{V}(s_{t+n})\Big)\end{aligned}$$

This rule is called TD(λ) learning

- $\lambda = 0$ recovers TD (or TD(0)).
- $\lambda \to 1$ recovers the Monte-Carlo method.

Part 2: Off-Policy Evaluation

OPE

Policy Evaluation

Given: A dataset $\mathcal{D}$ of trajectories trajectories $\tau_1, \dots, \tau_n$, sampled from behavior policy $\pi_{\rm b}$:

$$\begin{aligned} \tau_1 &= (s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_T^{(1)}, a_T^{(1)}, r_T^{(1)}) \\ &\vdots \qquad\qquad\qquad \vdots \\ \tau_n &= (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{T_n}^{(n)}, a_{T_n}^{(n)}, r_{T_n}^{(n)}) \end{aligned}$$

where

$$a_t^{(i)} \sim \boldsymbol{\pi}_{\rm b}(\cdot|s_t^{(i)}), \quad r_t^{(i)} \sim R(s_t^{(i)}, a_t^{(i)}), \quad s_{t+1}^{(i)} \sim P(\cdot|s_t^{(i)}, a_t^{(i)})$$

Goal: Derive (point) estimate, and possibly confidence intervals, for value of target policy π $(\neq \pi_{\rm b})$.

Each trajectory could be even sampled from a different behavior policy.

OPE Assumptions

The main challenge of OPE is mismatch of distributions π_{b} and π

Coverage Assumption

For all $s \in \mathcal{S}$, if $\pi(a|s) > 0$ then $\pi_{\mathrm{b}}(a|s) > 0$

Implication: π is absolutely continuous with respect to π_{b} (thus a.k.a. Absolute Continuity Assumption).

A Model-Based Method

Known Model

If MDP M known,

Then, $\quad V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$

for any $\pi \in \Pi^{\mathsf{SR}}$.

Idea: Estimate P and R via $\mathcal{D}$ and apply the certainty equivalence principle.

For simplicity, for now assume that $\mathcal{D}$ contains only one trajectory:

$$\mathcal{D} = \{(s_t, a_t, r_t), t = 1, \ldots, n\}$$

where:

$$a_t \sim \boldsymbol{\pi}_{\mathrm{b}}(\cdot|_t), \quad r_t \sim R(s_t, a_t), \quad s_{t+1} \sim P(\cdot|s_t, a_t)$$

A Model-Based Solution (I)

Idea: Estimate P and R via $\mathcal{D}$ and apply the certainty equivalence principle.

Introduce counts: For all (s, a, s')

$$N(s,a,s') = \sum_{t=1}^{n-1} \mathbb{I}\{s_t = s, a_t = a, s_{t+1} = s'\} \quad \text{and} \quad N(s,a) = \sum_{s' \in \mathcal{S}} N(s,a,s')$$

Smoothed Estimator for P and R:

$$\widehat{P}(s'|s,a) = \frac{N(s,a,s') + \alpha}{N(s,a) + \alpha S}, \qquad \widehat{R}(s,a) = \frac{\alpha + \sum_{t=1}^{n-1} r_t \mathbb{I}\{s_t = s, a_t = a\}}{\alpha + N(s,a)}$$

with $\alpha > 0$ an arbitrary smoothing parameter.

For any (s, a), if $\pi_{\mathrm{b}}(a|s) > 0$, then

$$\widehat{P}(\cdot|s,a) \to_{N(s)\to\infty} P(\cdot|s,a) \quad \text{and} \quad \widehat{R}(s,a) \to_{N(s)\to\infty} R(s,a), \qquad \text{almost surely.}$$

A Model-Based Solution (II)

Smoothed Estimator for P and R:

$$\widehat{P}(s'|s,a) = \frac{N(s,a,s') + \alpha}{N(s,a) + \alpha S}, \qquad \widehat{R}(s,a) = \frac{\alpha + \sum_{t=1}^{n-1} r_t \mathbb{I}\{s_t = s, a_t = a\}}{\alpha + N(s,a)}$$

$\Longrightarrow$ Build the empirical MDP $\widehat{M} = (\mathcal{S}, \mathcal{A}, \widehat{P}, \widehat{R}, \gamma)$.

Then, the following is an estimate for V^π:

$$\widehat{V}^\pi = (I - \gamma \widehat{P}^\pi)^{-1} \widehat{r}^\pi$$

with

$$\widehat{P}^\pi_{s,s'} = \sum_{a \in \mathcal{A}} \pi(a|s) \widehat{P}(s'|s,a) \qquad \text{and} \qquad \widehat{r}^\pi(s) = \sum_{a \in \mathcal{A}} \pi(a|s) \widehat{R}(s,a)$$

A Model-Based Solution (III)

Theorem

Under the coverage assumption and that all states are visited infinitely often under π_{b}, $\widehat{V}^{\pi}$ *converges to* V^{π} *almost surely:*

$$\mathbb{P}\Big(\lim_{n\to\infty}\widehat{V}^{\pi}=V^{\pi}\Big)=1$$

- In other words, if π_{b} is exploratory enough and the coverage assumption holds, $\widehat{V}^{\pi}$ converges to V^{π}.
- We can use concentration inequalities (e.g., Hoeffding's) to derive confidence interval(s) for V^{π}.

Model-Free Methods

Importance Sampling: Basic Facts

Consider two distributions P and Q defined on $\mathcal{X}$, with $P \ll Q$.

$$\mathbb{E}_{x\sim P}[f(x)] = \int_x f(x)P(x)\mathrm{d}x = \int_x f(x)Q(x) \underbrace{\frac{P(x)}{Q(x)}}_{\text{importance weight}} \mathrm{d}x = \mathbb{E}_{x\sim Q}\left[\frac{P(x)}{Q(x)}f(x)\right]$$

Note that importance weight $\frac{P(x)}{Q(x)}$ is well-defined due to $P \ll Q$.

Given are samples $X_i \sim Q, i = 1, \dots, n$:

- Importance weight estimator of $\mathbb{E}_{x\sim P}[f(x)]$:

$$\widehat{f}_{\mathsf{IS}} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{P(X_i)}{Q(X_i)}$$

- Importance weight estimator of $\mathbb{E}_{x\sim P}[f(x)]$:

$$\widehat{f}_{\mathsf{wIS}} = \frac{1}{\sum_{i=1}^{n}\frac{P(X_i)}{Q(X_i)}}\sum_{i=1}^{n} f(X_i)\frac{P(X_i)}{Q(X_i)}$$

Contrast these to $\widehat{f} = \frac{1}{n}\sum_{i=1}^{n} f(X_i)$ built using $X_i \sim P, i = 1, \dots, n$.

Importance Weight Estimators: Properties

Lemma

$\widehat{f}_{\mathsf{IS}}$ is consistent and unbiased.

Proof. Consistency follows from the SLLN. Unbiased since

$$\mathbb{E}[\widehat{f}_{\mathsf{IS}}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_Q[f(X_i)\tfrac{P(X_i)}{Q(X_i)}] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_P[f(X_i)] = \mathbb{E}_P[f(X)]$$

Lemma

$\widehat{f}_{\mathsf{wIS}}$ is consistent and biased.

Proof. To prove consistency, observe that by the SLLN, $\frac{1}{n}\sum_{i=1}^{n}\frac{P(X_i)}{Q(X_i)}$ converges to 1 w.p. 1 (since $X_i \sim Q$) and $\frac{1}{n}\sum_{i=1}^{n} f(X_i)\frac{P(X_i)}{Q(X_i)}$ converges to $\mathbb{E}_P[f(X)]$ w.p. 1. Showing biased via counter example: taking $X_1 = \ldots = X_n$,

$$\widehat{f}_{\mathsf{wIS}} = \frac{1}{\sum_{i=1}^{n}\frac{P(X_i)}{Q(X_i)}}\sum_{i=1}^{n} f(X_i)\frac{P(X_i)}{Q(X_i)} = f(X_1)$$

Hence, $\mathbb{E}[\widehat{f}_{\mathsf{wIS}}] = \mathbb{E}[f(X_1)] \neq \mathbb{E}[f(X_1)\frac{P(X_1)}{Q(X_1)}]$.

Importance Sampling Estimator for OPE

Consider a trajectory $\tau = (s_1, a_1, r_1, \dots, s_T, a_T, r_T)$ (with $s_1 = s$) sampled under $\pi_{\rm b}$.

Define t-step importance weight of τ as:

$$\rho_{1:t} = \prod_{t'=1}^{t} \frac{\pi(a_{t'}|s_{t'})}{\pi_{\rm b}(a_{t'}|s_{t'})}$$

In fact, $\frac{\mathbb{P}(\tau|\pi)}{\mathbb{P}(\tau|\pi_{\rm b})} = \rho_{1:T}$.
Importance sampling estimator of $V^\pi(s)$ built using τ:

$$\widehat{V}^{\boldsymbol{\pi}}_{\sf IS}(s;\tau) = \frac{\mathbb{P}(\tau|\pi)}{\mathbb{P}(\tau|\pi_{\rm b})} \sum_{t=1}^{T} \gamma^{t-1} r_t = \boldsymbol{\rho_{1:T}} \sum_{t=1}^{T} \gamma^{t-1} r_t$$

Contrast it with $\widehat{V}^{\pi_{\rm b}}(s) = \sum_{t=1}^{T} \gamma^{t-1} r_t$ built using τ.

Importance Sampling Estimator for OPE

Given a dataset $\mathcal{D}$ of n trajectories $\tau_1, \dots, \tau_n$:

$$\tau_1 = (s_1^{(1)}, a_1^{(1)}, r_1^{(1)}, \dots, s_T^{(1)}, a_T^{(1)}, r_T^{(1)})$$

$$\vdots \qquad\qquad \vdots$$

$$\tau_n = (s_1^{(n)}, a_1^{(n)}, r_1^{(n)}, \dots, s_{T_n}^{(n)}, a_{T_n}^{(n)}, r_{T_n}^{(n)})$$

all staring in s (i.e., $s_1^{(1)} = \dots = s_1^{(n)} = s$).

- Importance sampling estimator of $V^\pi(s)$ built using $\mathcal{D}$:

$$\widehat{V}^\pi_{\mathsf{IS}}(s; \mathcal{D}) = \frac{1}{n} \sum_{i=1}^n \widehat{V}^\pi_{\mathsf{IS}}(s; \tau_i) = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\rho}^{(i)}_{1:T_i} \sum_{t=1}^T \gamma^{t-1} r_t^i$$

  (unbiased, but typically with high variance)

- Weighted importance sampling estimator of $V^\pi(s)$ built using $\mathcal{D}$:

$$\widehat{V}^\pi_{\mathsf{wIS}}(s; \mathcal{D}) = \frac{\sum_{i=1}^n \boldsymbol{\rho}^{(i)}_{1:T_i} \sum_{t=1}^T \gamma^{t-1} r_t^i}{\sum_{i=1}^n \boldsymbol{\rho}^{(i)}_{1:T_i}}$$

  (slightly biased, but with lower variance)

Next lecture: Further on OPE + algorithms for OPO!