model-free methods

goal: approximate optimal value & policy without P & c, i.e. from data

Bertsekas & Tsitsiklis Ch 5

- so far, we've assumed given an MDP $(X, \mathcal{U}, P, c)$ so that

$$\min_u E[c(x,u)] \text{ s.t. } x^+ \sim P(x,u)$$

can be solved exactly using value / policy iteration

→ what would you do if you weren't given P? c?

- if not given P, could observe a large number of controlled trajectories / rollouts and form an estimate of P
- if not given c, at least along traj's / rollouts, the problem becomes [illegible]

many EPs/rollouts, the
problem becomes unsupervised
⇝ not clear what to do!
↳ though possible in principle,
can be inefficient in practice
⇝ we'll develop alternative methods

Monte Carlo estimation

- let $v : \Omega \to \mathbb{R}$ be a random variable,
  $$\bar{v}_N = \frac{1}{N} \sum_{n=1}^{N} v_n$$ is termed
  the <u>sample mean</u> of dataset $\{v_n\}_{n=1}^{N}$

→ show that sample mean can be
computed recursively

- $\bar{v}_{N+1} = \bar{v}_N + \frac{1}{N+1}(v_{N+1} - \bar{v}_N)$

- so long as samples $\{v_n\}$ are
  iid (independent, identically distributed)
  then $E[\bar{v}_N] = E[v]$,
  so $\bar{v}_N$ is unbiased

so $\bar{v}_N$ is <u>unbiased</u>

- now given policy $\mu : X \to \Delta(\mathcal{U})$, if we generate traj's

$$\{(x_n, u_n) : [0,t] \to X \times \mathcal{U}\}_{n=1}^{N}$$

from initial state $x_n(0) = \xi$, then we can estimate

$$v^{\mu}(\xi) = E[c \mid x(0) = \xi]$$
$$\simeq \frac{1}{N} \sum_{n=1}^{N} c(x_n, u_n) = \bar{v}_N^{\mu}(\xi)$$

– note that this sample mean could be computed iteratively

$$\bar{v}_{N+1}^{\mu}(\xi) = \bar{v}_N^{\mu} + \frac{1}{N+1} \cdot \left(c_{N+1} - \bar{v}_N^{\mu}\right)$$

* we will see many variations on this form of iterative update to an estimate of e.g. value function

---

$$\boxed{\text{temporal differences (TD)}}$$

- let's return to the iterative estimate

$$v^+(x_0) = v(x_0) + \alpha \cdot \left(c(x, u) - v(x_0) \right)$$

- substituting out cost

$$c(x, u) = \sum_{t=0}^{\infty} \gamma^t \cdot \mathcal{L}(x_t, u_t)$$

and rearranging slightly:

we've added nothing (zero)

$$
\begin{aligned}
v^+(x_0) = v(x_0) \\
&+ \alpha \cdot \Big[\gamma^0 \cdot \left(\mathcal{L}(x_0, u_0) + \gamma \cdot v(x_1) - v(x_0) \right) \\
&\quad + \gamma^1 \cdot \left(\mathcal{L}(x_1, u_1) + \gamma \cdot v(x_2) - v(x_1) \right) \\
&\quad + \dots \\
&\quad + \gamma^t \cdot \left(\mathcal{L}(x_t, u_t) + \gamma \cdot v(x_{t+1}) - v(x_t) \right) \\
&\quad + \dots \Big]
\end{aligned}
$$

or, equivalently,

$$v^+(x_0) = v(x_0) + \alpha \cdot \sum_{t=0}^{\infty} \gamma^t \cdot \underbrace{\left(\mathcal{L}(x_t, u_t) + \gamma \cdot v(x_{t+1}) - v(x_t) \right)}_{\text{temporal differences (TD)}}$$

- TD $d_t = \mathcal{L}(x_t, u_t) + \gamma \cdot v(x_{t+1}) - v(x_t)$

– TD $d_t = \mathcal{L}(x_t, u_t) + \gamma \cdot v(x_{t+1}) - v(x_t)$

represents difference between two estimates of same quantity:

$$v(x_t) \quad \text{vs} \quad \mathcal{L}(x_t, u_t) + v(x_{t+1})$$

– noting that d_t becomes available at time t, we are tempted to update $v(x_t) = v(x_t) + \alpha \cdot d_t$ as soon as d_t becomes available

- more generally, the TD update above could be derived from single-sample estimate of Bellman equation

$$v^\mu(x) = E\left[\mathcal{L}(x, u) + \gamma \cdot v^\mu(x^+)\right]$$

– this suggests multi-sample variant

$$v^\mu(x_0) = E\left[\sum_{s=0}^{t} \gamma^t \cdot \mathcal{L}(x_s, u_s) + \gamma^{t+1} \cdot v^\mu(x_{t+1})\right]$$

– and, finally, weighted average: fix $\lambda \in (0,1)$ and compute

$$v^\mu(x_0) = (1-\lambda) \cdot E\left[\sum_{t=0}^{\infty} \lambda^t \left(\sum_{s=0}^{t} \gamma^s \cdot \mathcal{L}(x_s, u_s) + \gamma^{t+1} \cdot v^\mu(x_{t+1})\right)\right]$$

$= \cdots$ (exchange order of Σ's,

$$= \cdots \quad \left(\text{exchange order of } \Sigma\text{'s, use } \lambda^t = (1-\lambda)\sum_{s=t}^{\infty} \lambda^s\right)$$

$$= E\left[\sum_{t=0}^{\infty} \lambda^t \cdot \gamma^t \cdot d_t\right] + v^\mu(x_0), \text{ where}$$

$$d_t = \mathcal{L}(x_t, u_t) + \gamma \cdot v^\mu(x_{t+1}) - v^\mu(x_t)$$

→ why is this equation obvious?
why is it nevertheless useful?

- obvious since Bellman tells us

$$v^\mu(x_0) - v^\mu(x_0) = E\left[\sum_{t=0}^{\infty} \lambda^t \cdot \gamma^t \cdot d_t\right] = 0$$

- useful since we can apply Monte Carlo estimation:

$$v^+(x_0) = v(x_0) + \alpha \cdot \sum_{t=0}^{\infty} \lambda^t \cdot \gamma^t \cdot d_t$$

- varying λ from 1 to 0 interpolates from Monte Carlo to TD(0), above, yielding TD(λ)

<u>fact</u>: if each state is visited by infinitely many traj's and $\alpha \to 0$, TD(λ) converges to v^μ in probability

infinitely many trips and $\alpha \to 0$,

TD(λ) converges to v^μ in probability

(proof follows contraction argument)

- there are other variations, eg including eligibility traces, that may improve performance in practice

---

WARNING

- preceding methods can asymptotically approximate v^μ, i.e. the value of a particular policy, μ
- to obtain a complete RL algorithm, need to switch from evaluating current policy to improving policy
- if policy improvement is performed with non-asymptotic $\tilde{v}^\mu$, termed optimistic policy iteration, algorithm can fail to converge!

algorithm can fail to converge!
(see sec 5.5 in B&T 96)

Q-learning

- given the value $v^\mu \in \mathbb{R}^X$ of the policy $\mu : X \to \Delta(\mathcal{U})$, consider the state-action quality

$$q^\mu : X \times \mathcal{U} \to \mathbb{R}$$

$$: (x,u) \mapsto \sum_{x^+ \in X} P(x^+ | x,u) \cdot \left(c(x,u) + \gamma \cdot v^\mu(x^+) \right)$$

i.e. the expected cost applying control u at state x, then applying μ

→ how does q^μ relate to v^μ?

– $v^\mu(x) = \min_{u \in \mathcal{U}} q^\mu(x,u)$

→ given q^μ, how would you improve μ?

– $\mu^+(x) = \arg\min_{u \in \mathcal{U}} q^\mu(x,u)$

– why define q^μ?

– why define q^μ?

* it will turn out that q^μ is easier to estimate from data:

– note that the optimal q^* satisfies the equation

$$q^*(x,u) = E\left[\mathcal{L}(x,u) + \gamma \cdot \min_{w \in \mathcal{U}} q^*(x,w)\right]$$

(where expectation is over x^+'s)

– applying Monte Carlo estimation, obtain q-learning algorithm:

$$q^+(x,u) = (1-\alpha) \cdot q(x,u) + \alpha \cdot \left(\mathcal{L}(x,u) + \gamma \cdot \min_{w \in \mathcal{U}} q(x,w) \right)$$

– looks similar to TD(0); generalizing to multi-step case non-obvious

fact: q-learning converges in probability