# RL with function approximation

$\Rightarrow$ $x_t, g_t$ $t \geq 0$ = state/cost realizations when MDP is operated with policy $\mu(\cdot)$ (stochastic)

$$x_{t+1} = f(x_t, \underbrace{\mu(x_t)}_{u_t}, w_t)$$
$$g_t = g(x_t, \underbrace{\mu(x_t)}_{u_t}, w_t)$$

we want to estimate

$V^\mu(x)$, $x \in X$ $\quad Q^\mu(x,u)$ $\quad x,u \in X \times U$ $\Rightarrow$ then use GPI

however: $|X| = n$ and $|U| = m$ are <u>too big</u> for an exact representation of $V^\mu(x)$ and $Q^\mu(x,u)$ $\to$ chess no. of board configurations $> 10^{100}$

$\Rightarrow$ use function approximation

$$V^\mu(x) \approx \hat{V}(x,\vartheta) \qquad Q^\mu(x,u) \approx \hat{Q}(x,u,\vartheta) \qquad \vartheta \in \mathbb{R}^q \text{ ("small")}$$

requirements: store $\vartheta$ and evaluate $\hat{V}(x,\vartheta)$ and $\hat{Q}(x,u,\vartheta)$ for given $x$ and $u$

<u>linear</u>: $\hat{V}(x,\vartheta) = \varphi(x)^T \vartheta$ $\qquad \varphi(x)$ $x \in X$ = basis function

$\hat{Q}(x,u,\vartheta) = \varphi(x,u)^T \vartheta$ $\qquad \varphi(x,u)$ $x,u \in X \times U$ "

$\to$ less modeling capabilities, strong convergence properties

tabular case as a particular case:

$$\vartheta \in \mathbb{R}^n \quad \varphi(i) = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \text{-ith} \quad \Rightarrow \quad \varphi(i)^T \vartheta = \vartheta_i = V^\mu(i)$$

mutatis mutandis for $Q^\mu(x,u)$

<u>non-linear</u>

e.g. $\hat{V}(x,\vartheta)$ and $\hat{Q}(x,u,\vartheta)$ = Neural Network

$\to$ convergence critical, strong modeling power

$V^\mu(x)$ to begin with ($Q^\mu(x,u)$ conceptually identical)

Approximation achieved via projection onto $\mathcal{M} = \{\hat{V}(x,\vartheta) : \vartheta \in \mathbb{R}^q\}$

$\Rightarrow$ Given $V(x)$, let

$$\Pi[V(x)] = \underset{\hat{V}\in\mathcal{M}}{\arg\min}\ \mathbb{E}\left[\left(V(x_t) - \hat{V}(x_t,\vartheta)\right)^2\right]$$

$\rightarrow$ expectation taken w.r.t. to the stationary distribution of $x_t$ when MDP operated with $\mu(\cdot)$

$$\| \qquad \hat{V}(x,\bar{\vartheta}) \text{ when } \bar{\vartheta} = \underset{\vartheta}{\arg\min} \sum_{i=1}^{M} \xi_i \left(V(i) - \hat{V}(i,\vartheta)\right)^2$$

$\hookrightarrow \xi_i = \mathbb{P}\{x_t = i\} \quad t \to \infty$

<u>DIRECT (BATCH) APPROACH</u>

goal: compute $\Pi[V^\mu(x)]$

$$\mathbb{E}\left[\left(V^\mu(x_t) - \hat{V}(x_t,\vartheta)\right)^2\right] \approx \frac{1}{T}\sum_{t=0}^{T-1}\left(V^\mu(x_t) - \hat{V}(x_t,\vartheta)\right)^2$$

$\uparrow$ available realization $x_1, x_2, \ldots$

$$V^\mu(x_\ell) \approx \sum_{\ell=t}^{T-1} \gamma^{\ell-t} g_\ell \approx \sum_{i=0}^{+\infty} \gamma^i g_{t+i}$$

$$\Rightarrow \bar{\vartheta} = \underset{\vartheta}{\arg\min}\ \frac{1}{T}\sum_{t=0}^{T-1}\left(\hat{V}(x_t,\vartheta) - \sum_{i=0}^{T-1-t}\gamma^i g_{t+i}\right)^2$$

▸ can be solved via a gradient method

$$\hat{\vartheta}_{k+1} = \hat{\vartheta}_k - \alpha \sum_{t=0}^{T-1} \nabla_\vartheta \hat{V}(x_t,\hat{\vartheta}_k)\left(\hat{V}(x_t,\hat{\vartheta}_k) - \sum_{i=0}^{T-1-t}\gamma^i g_{t+i}\right)$$

$\rightsquigarrow$ stochastic gradient can be used

$\uparrow$ to be modulated

stochastic gradient:

$$\hat{\vartheta}_{k+1} = \hat{\vartheta}_k - \alpha \nabla_\vartheta \hat{V}(x_k,\hat{\vartheta}_k)\left(\hat{V}(x_k,\vartheta_k) - \sum_{i=0}^{T-1-k}\gamma^i g_{k+i}\right)$$

$\rightarrow$ the whole batch is still needed $\rightarrow$ use short T and change batch memory

$\sum_{\ell=0}^{T-1}\sum_{t=0}^{\ell}$ and then $\ell \to t \quad t \to \ell$

<u>OBS</u>: we can write

$$\hat{\vartheta}_{k+1} = \hat{\vartheta}_k - \alpha\left(\sum_{t=0}^{T-1}\nabla_\vartheta\hat{V}(x_t,\hat{\vartheta}_k)\hat{V}(x_t,\hat{\vartheta}_k) - \sum_{t=0}^{T-1}\sum_{\ell=t}^{T-1}\nabla_\vartheta\hat{V}(x_t,\hat{\vartheta}_k)\gamma^{\ell-t}g_\ell\right)$$

$$= \hat{\vartheta}_k - \alpha\left(\sum_{t=0}^{T-1}\nabla_\vartheta\hat{V}(x_t,\hat{\vartheta}_k)\hat{V}(x_t,\hat{\vartheta}_k) - \sum_{t=0}^{T-1}\sum_{\ell=0}^{t}\nabla_\vartheta\hat{V}(x_\ell,\hat{\vartheta}_k)\gamma^{t-\ell}g_t\right)$$

incremental update: can be used with very a long batch

stochastic gradient:

$$\hat{\vartheta}_{k+1} = \hat{\vartheta}_k - \alpha\left(\nabla_\vartheta\hat{V}(x_k,\hat{\vartheta}_k)\hat{V}(x_k,\hat{\vartheta}_k) - \underbrace{\sum_{\ell=0}^{k}\nabla_\vartheta\hat{V}(x_\ell,\hat{\vartheta}_k)\gamma^{k-\ell}}_{\text{can be incrementally computed}} g_k\right)$$

$\hat{\vartheta}_\ell \to \hat{\vartheta}_\ell$

INDIRECT APPROACHES $\Rightarrow$ projected Bellman equation

$$V^\mu(x): \quad V^\mu(x) = T^\mu[V^\mu(x)] = \mathbb{E}\left[g(x,\mu(x),w) + \gamma V^\mu(f(x,\mu(x),w))\right]$$

We consider a projection of this equation over the space $\mathcal{M}$

$$\hat{V}(x,\bar\vartheta) = \underbrace{\Pi}_{\text{projection over } \mathcal{M}}\left[T^\mu[\hat{V}(x,\bar\vartheta)]\right] \qquad \hat{V}(x,\vartheta) \in \mathcal{M} \quad T^\mu[\hat{V}(x,\vartheta)] \notin \mathcal{M}$$

more explicitely

w.r.t. $x_t$ w.r.t. $w_t$

$$\hat{V}(x,\bar\vartheta) = \underset{\hat{V}(x,\vartheta)\in\mathcal{M}}{\arg\min}\ \mathbb{E}_{x_t}\left[\left(\hat{V}(x_t,\vartheta) - \mathbb{E}_{w_t}\left[g(x_t,\mu(x_t),w_t) + \gamma \hat{V}(f(x_t,\mu(x_t),w_t),\bar\vartheta)\right]\right)^2\right]$$

$\Updownarrow$ if $\hat{V}(\cdot,\vartheta)$ one to one ... if $\hat{V}(x,\vartheta) = \varphi(x)^T\vartheta$ then $\varphi(x)$ linear ind.

$[\varphi(1) \dots \varphi(n)]$ full rank

$$\bar\vartheta = \underset{\vartheta}{\arg\min}\ \mathbb{E}\left[\left(\hat{V}(x_t,\vartheta) - \left(g(x_t,\mu(x_t),w_t) + \gamma \hat{V}(f(x_t,\mu(x_t),w_t),\bar\vartheta)\right)\right)^2\right]$$

assume the projected equation has one and only one solution (under certain condition, it does $\to$ see later)

then $\bar\vartheta :=$ solution to $\nabla_\vartheta\left(\mathbb{E}\left[(\ ---\ )^2\right]\right) = 0$ i.e. it must be

$$\mathbb{E}\left[\nabla_\vartheta \hat{V}(x_t,\bar\vartheta)\cdot\left(\hat{V}(x_t,\bar\vartheta) - \left(g(x_t,\mu(x_t),w_t) + \gamma\hat{V}(f(x_t,\mu(x_t),w_t),\bar\vartheta)\right)\right)\right] = 0$$

can be written as a <u>fixed point problem</u>: find $\bar\vartheta$ such that

$$\bar\vartheta = \bar\vartheta - \mathbb{E}\left[\nabla_\vartheta \hat{V}(x_t,\bar\vartheta)\cdot\left(\hat{V}(x_t,\bar\vartheta) - \left(g(x_t,\mu(x_t),w_t) + \gamma\hat{V}(f(x_t,\mu(x_t),w_t),\bar\vartheta)\right)\right)\right]$$

can be obtained via stochastic approximation + bootstrapping

$x_t, g_t$ $t=0,1,\dots$ realization of state and cost

$\hat\vartheta_t =$ estimate of $\bar\vartheta$ at time $t$

given $\hat\vartheta_t$, new target become

bootstrapped stochastic realization of the right-hand side of

$$\hat\vartheta_t + \nabla_\vartheta \hat{V}(x_t,\hat\vartheta_t)\left(g_t + \gamma\hat{V}(x_{t+1},\hat\vartheta_t) - \hat{V}(x_t,\hat\vartheta_t)\right)$$

and direction from $\hat{\vartheta}_t$ towards target:

$$\nabla_\vartheta \hat{V}(x_t, \hat{\vartheta}_t)\underbrace{\left(g_t + \gamma \hat{V}(x_{t+1}, \hat{\vartheta}_t) - \hat{V}(x_t, \hat{\vartheta}_t)\right)}_{\text{Temporal Difference}}$$

TD(0) algorithm:

$$\begin{cases} \delta_t = g_t + \gamma \hat{V}(x_{t+1}, \hat{\vartheta}_t) - \hat{V}(x_t, \hat{\vartheta}_t) \\ \hat{\vartheta}_{t+1} = \hat{\vartheta}_t + \alpha_t \delta_t \cdot \nabla_\vartheta \hat{V}(x_t, \hat{\vartheta}_t) \end{cases}$$

In the linear case $\hat{V}(x, \vartheta) = \varphi(x)^T \vartheta \Rightarrow \nabla_\vartheta \hat{V}(x, \vartheta) = \varphi(x)$

$$\begin{cases} \delta_t = g_t + \gamma \varphi(x_{t+1})^T \hat{\vartheta}_t - \varphi(x_t)^T \hat{\vartheta}_t \\ \hat{\vartheta}_{t+1} = \hat{\vartheta}_t + \alpha_t \delta_t \varphi(x_t) \end{cases}$$

when $\varphi(i) = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ - i-th

$\varphi(x_t) = \begin{bmatrix} \mathbb{1}_{x_t = 1} \\ \vdots \\ \mathbb{1}_{x_t = M} \end{bmatrix} \Rightarrow$ back to the tabular case!

In the linear case, projected Bellman equation has indeed one and only one solution! $\Rightarrow$ convergence of TD(0) well-established

$$\varphi(x)^T \bar{\vartheta} = \underset{\downarrow}{\Pi}\left[ T^\mu \left( \varphi(x)^T \bar{\vartheta} \right) \right]$$

idea: $\Pi\left[T^\mu(\cdot)\right]$ is a contraction

projection onto a linear subspace = orthogonal projection

$$\Pi\left[V(x)\right] = \varphi(x)^T \cdot \arg\min_\vartheta \mathbb{E}\left[ \left( \varphi(x_t)^T \vartheta - V(x_t) \right)^2 \right] =$$

$$= \varphi(x)^T \cdot \underbrace{\mathbb{E}\left[ \varphi(x_t) \varphi(x_t)^T \right]^{-1}}_{\substack{\text{exist and is unique if } \varphi(x) \text{ are linear independent } + \\ \text{some assumptions on the stationary distribution of } x_t}} \cdot \mathbb{E}\left[ \varphi(x_t) V(x_t) \right]$$

linear operator $\downarrow$ linear in $V(x)$

Let $\|V(x)\|_S^2 = \underset{\substack{\uparrow \\ \text{stationary distribution of } x_t}}{\mathbb{E}}\left[ V(x_t)^2 \right]$ = same norm (and scalar product) used to define the projection $\Pi$

Then, given $V_1(x)$ and $V_2(x)$ we have:

$$\left\| \Pi\left[V_1(x)\right] - \Pi\left[V_2(x)\right] \right\|_S^2 \underset{\text{linearity}}{=} \left\| \Pi\left[ V_1(x) - V_2(x) \right] \right\|_S^2 \leq$$

$$\leq \left\| \Pi\left[V_1(x)-V_2(x)\right]\right\|_S^2 + \left\| V_1(x)-V_2(x) - \Pi\left[V_1(x)-V_2(x)\right]\right\|_S^2$$

orthogonal $\Rightarrow$ use Pytagora's Th.

$$= \left\| V_1(x) - V_2(x)\right\|_S^2$$

Hence, $\left\| \Pi[V_1(x)] - \Pi[V_2(x)]\right\|_S \leq \left\| V_1(x) - V_2(x)\right\|_S$ and

$$\left\| \Pi\left[T^\mu(V_1(x))\right] - \Pi\left[T^\mu(V_2(x))\right]\right\|_S \leq \left\| T^\mu(V_1(x)) - T^\mu(V_2(x))\right\|_S$$

$$\leq \beta \left\| V_1(x) - V_2(x)\right\|_S$$

$\beta < 1$

for issue $T^\mu(\cdot)$ contractive w.r.t. to the max norm, also for $\|\cdot\|_S$ under some condition on the stationary distribution + linear independence of $\varphi(x)$

contraction!

$\Downarrow$

projected Bellman eq always has a unique solution and convergence of TD(0) is guaranteed

In the linear case, alternatives to solve approximately the projected Bellman equation exist!

$$\bar{\vartheta} = \arg\min_{\vartheta} \mathbb{E}\left[\left(\varphi(x_t)^T\vartheta - \left(g(x_t,\mu(x_t),w_t) + \gamma\,\varphi(f(x_t,\mu(x_t),w_t))^T\bar{\vartheta}\right)^2\right]$$

$$\approx \arg\min_{\vartheta} \underbrace{\frac{1}{T}\sum_{t=0}^{T-1}}\left(\varphi(x_t)^T\vartheta - g_t - \gamma\,\varphi(x_{t+1})^T\bar{\vartheta}\right)^2$$

Least Squares Temporal Difference

data driven approximation, over a realization $x_t, g_t$ $t=0,1,\ldots$

LSTD: solve the Least Square problem (quadratic in $\vartheta$)

$$\hat{\vartheta}_T: \frac{1}{T}\sum_{t=0}^{T-1}\varphi(x_t)\cdot\left[\left(\varphi(x_t)^T - \gamma\,\varphi(x_{t+1})^T\right)\bar{\vartheta} - g_t\right] = 0$$

$$\Rightarrow \hat{\vartheta}_T = \left[\sum_{t=0}^{T-1}\varphi(x_t)\left(\varphi(x_t)^T - \gamma\,\varphi(x_{t+1})\right)\right]^{-1} \times \sum_{t=0}^{T-1}\varphi(x_t)\cdot g_t$$

recursive implementation are possible so as to avoid memory saturation and computational issues easiest

$$\begin{cases} C_t = C_{t-1} + \varphi(x_t)\left(\varphi(x_t)^T - \gamma\varphi(x_{t+1})\right) \\ d_t = d_{t-1} + \varphi(x_t)\, g_t \\ \hat{\vartheta}_{t+1} = C_t^{-1} \cdot d_t \end{cases}$$

$\rightsquigarrow$ refined version that avoid inversion exist

**LSPE** : use value iteration for the projection operator

$$\hat{\vartheta}_{t+1} = \arg\min_{\vartheta} \mathbb{E}\left[\left(\varphi(x_t)^T\vartheta - \left(g(x_t, \mu(x_t), w_t) + \gamma\varphi\left(f(x_t,\mu(x_t),w_t)\right)^T\hat{\vartheta}_t\right)\right)^2\right]$$

$$\simeq \arg\min_{\vartheta} \frac{1}{t+1}\sum_{j=0}^{t}\left(\varphi(x_j)^T\vartheta - g_j - \gamma\varphi(x_{j+1})^T\hat{\vartheta}_t\right)^2$$

$$= \left[\sum_{j=0}^{t}\varphi(x_j)\varphi(x_j)^T\right]^{-1} \cdot \sum_{j=0}^{t}\varphi(x_j)\left[g_j + \gamma\varphi(x_{j+1})^T\hat{\vartheta}_t \;\; \hat{} \; \pm\varphi(x_j)^T\hat{\vartheta}_t\right]$$

$$= \hat{\vartheta}_t + \left[\sum_{j=0}^{t}\varphi(x_j)\varphi(x_j)^T\right]^{-1} \cdot \sum_{j=0}^{t}\varphi(x_j)\underbrace{\left(g_j + \gamma\varphi(x_{j+1})\hat{\vartheta}_t - \varphi(x_j)^T\hat{\vartheta}_t\right)}_{TD}$$

Least Square Policy Evolution

possibly changed to $\hat{\vartheta}_j$ to form an incremental implementati

TD($\lambda$), LSTD($\lambda$), LSPE($\lambda$) are also possible

$\rightsquigarrow$ repeat everything with $(1-\lambda)\sum_{k=1}^{+\infty}\lambda^{k-1}(T^\mu)^k$ in place of $T^\mu$ $\rightsquigarrow$ cumbersome calculations, get easy to implement via eligibility traces

TD($\lambda$):

$$\begin{cases} \delta_t = g_t + \gamma\hat{V}(x_{t+1}, \hat{\vartheta}_t) - \hat{V}(x_t, \hat{\vartheta}_t) \\ z_t = \nabla_\vartheta \hat{V}(x_t, \hat{\vartheta}_t) + \gamma\lambda\, z_{t-1} \\ \hat{\vartheta}_{t+1} = \hat{\vartheta}_t + \alpha_t\,\delta_t\, z_t \end{cases}$$

LSTD($\lambda$):

$$z_t = \varphi(x_t) + \gamma\lambda z_{t-1}$$

$$\hat{\vartheta}_{t+1} = \left[\sum_{j=0}^{t} z_t \left(\varphi(x_j)^T - \gamma\varphi(x_{j+1})^T\right)\right]^{-1} \cdot \sum_{j=0}^{t} z_j \cdot g_j$$

LSPE($\lambda$):

$$z_t = \varphi(x_t) + \gamma\lambda z_{t-1}$$

$$\hat{\vartheta}_{t+1} = \hat{\vartheta}_t + \left[\sum_{j=0}^{t} \varphi(x_j)\varphi(x_j)^T\right]^{-1} \cdot \sum_{j=0}^{t} z_j \left(g_j + \gamma\varphi(x_{j+1})\hat{\vartheta}_t - \varphi(x_j)^T\hat{\vartheta}_t\right)$$

Everything can be repeated for $\hat{Q}(x,u,\vartheta)$ without conceptual changes

$$\hat{V}(x_t,\vartheta) \longrightarrow \hat{Q}(x_t, u_t, \vartheta)$$

$$\hat{V}(x_{t+1},\vartheta) \longrightarrow \hat{Q}(x_{t+1}, \mu(x_{t+1}), \vartheta) = \hat{Q}(x_{t+1}, u_{t+1}, \vartheta) \quad \text{(on-policy)}$$

$$\varphi(x_t) \longrightarrow \varphi(x_t, u_t)$$

$$\varphi(x_{t+1}) \longrightarrow \varphi(x_{t+1}, \mu(x_{t+1})) = \varphi(x_{t+1}, u_{t+1}) \quad \text{(on policy)}$$

$\Rightarrow$ estimating $\hat{Q}(x,u,\vartheta)$ for model-free policy improvement