LSTD, LSPE and TD(α) Methods

LSTD (Least Square Temporal Difference) method

$$r_k = C_k^{-1} d_k \quad , \quad k=0,1,2,\cdots$$

- r_k can be obtained by solving $C_k r_k = d_k$ using a batch of $k+1$ samples and (30).
- Alternatively, LSTD can be written as

$$C_k r_k - d_k = \frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\, g_{k,t} = 0 \qquad \text{——— (33)}$$

where the scalar $g_{k,t}$ is the so-called temporal difference (TD for short), associated with r_k and transition (i_t, i_{t+1}), given by

$$g_{k,t} = \phi(i_t)' r_k - \alpha\, \phi(i_{t+1})' r_k - g(i_t, i_{t+1}) . \qquad \text{——— (34)}$$

- TD may be viewed as a sample of a residual term arising in the projected Bellman equation, i.e.,

$$\tilde{J}(i_t; r_k) = g_{k,t} + g(i_t, i_{t+1}) + \alpha\, \tilde{J}(i_{t+1}; r_k)$$

associated with r_k and transition (i_t, i_{t+1}).

LSPE (Least Square Policy Evaluation) method

$$r_{k+1} = r_k - \gamma G_k (C_k r_k - d_k) \quad , \quad k=0,1,2,\cdots \qquad \text{——— (35)}$$

where γ is the stepsize and G_k is a scaling matrix that converges to some G such that spectral radius of $I - \gamma G C$ is smaller than 1.

- One possibility is to choose $\gamma = 1$ and

$$G_k = \left(\frac{1}{k+1} \sum_{t=0}^{k} \phi(i_t)\phi(i_t)' \right)^{-1}$$

where one can show that $G_k \to G$ as $C_k \to C$ and $d_k \to d$ due to the

law of large numbers argument.

- Alternatively, LSPE can be written as

$$r_{k+1} = r_k - \frac{\gamma}{k+1} G_k \sum_{t=0}^{k} \phi(i_t) q_{k,t} \qquad (36)$$

TD(0) method

- May viewed as a single sample approximation of the simulation-based general PVI iteration method with $G = I$

$$r_{k+1} = r_k - \gamma (C_k r_k - d_k)$$

$$= r_k - \frac{\gamma}{k+1} \sum_{t=0}^{k} \phi(i_t) q_{k,t} \qquad (37)$$

$$\approx r_k + \gamma_k \phi(i_k) q_{k,k} \quad \text{with } \gamma_k = \frac{\gamma}{k+1} \to 0$$

- TD(0) is much slower than LSPE or (37) but requires much less overhead

- TD(0) can be also viewed as a stochastic approximation method of the following fixed point equation

$$r = r - \gamma (Cr - d) \quad \Leftarrow \quad Cr = d.$$

Multi-step PVI Methods

- Bellman's eq. for a stationary policy μ

$$J_\mu = T_\mu J_\mu = g_\mu + \alpha P_\mu J_\mu$$

- Replace T_μ with a multi-step operator that has the same fixed point of Bellman's equation. For example,

$$T_\mu^\ell, \quad \ell > 1$$

or $$T_\mu^{(\lambda)} = (1-\lambda) \sum_{\ell=0}^{\infty} \lambda^\ell T_\mu^{\ell+1}, \quad \lambda \in (0,1)$$

Why?

$$J_\mu = T_\mu J_\mu$$
$$\lambda J_\mu = \lambda T_\mu^2 J_\mu$$
$$\lambda^2 J_\mu = \lambda^2 T_\mu^3 J_\mu$$
$$\vdots$$
$$\lambda^\ell J_\mu = \lambda^\ell T_\mu^{\ell+1} J_\mu$$
$$\vdots$$
$$\Rightarrow (1+\lambda+\cdots) J_\mu = \sum_{\ell=0}^{\infty} \lambda^\ell T_\mu^{\ell+1} J_\mu$$
$$\Rightarrow J_\mu = (1-\lambda) \sum_{\ell=0}^{\infty} \lambda^\ell T_\mu^{\ell+1} J_\mu$$

$$J_\mu = T_\mu^{(\lambda)} J_\mu$$

where $$T_\mu^{(\lambda)} = (1-\lambda) \sum_{\ell=0}^{\infty} \lambda^\ell T_\mu^{\ell+1}$$

" λ- weighted multi-step Bellman Equation"

(38)

- Note that

$$J_\mu = T_\mu^{(\lambda)} J_\mu = g_\mu^{(\lambda)} + \alpha P_\mu^{(\lambda)} J_\mu$$

where $g_\mu^{(\lambda)} = \sum_{\ell=0}^{\infty} \alpha^\ell \lambda^\ell P_\mu^\ell g_\mu^\ell = (I - \alpha\lambda P_\mu)^{-1} g_\mu$

$$P_\mu^{(\lambda)} = (1-\lambda) \sum_{\ell=0}^{\infty} \alpha^\ell \lambda^\ell P_\mu^{\ell+1} \qquad \text{——— (39)}$$

- Projected Bellman equation for $T_\mu^{(\lambda)}$

$$\Phi r = \Pi T_\mu^{(\lambda)} \Phi r$$

or

$$C^{(\lambda)} r = d^{(\lambda)}$$

where $C^{(\lambda)} = \Phi' M (I - \alpha P_\mu^{(\lambda)}) \Phi$, $d^{(\lambda)} = \Phi' M g_\mu^{(\lambda)}$.

Proposition 15.

Let Assumptions 1 and 2 hold, and let Π be the projection w.r.t. the weighted Euclidean norm $\|\cdot\|_\xi^2$ where ξ is the steady-state probability vector of the Markov chain corresponding to the given policy μ. Then,

(a) The mappings $T_\mu^{(\lambda)}$ and $\Pi T_\mu^{(\lambda)}$ corresponding to μ are contractions of modulus

$$\alpha_\lambda = \frac{\alpha(1-\lambda)}{1-\alpha\lambda}$$

w.r.t. $\|\cdot\|_\xi^2$.

(b) We have

$$\|J_\mu - \Phi r_\lambda^*\|_\xi^2 \le \frac{1}{\sqrt{1-\alpha_\lambda^2}} \|J_\mu - \Pi J_\mu\|_\xi^2$$

where Φr_λ^* is the unique fixed point of $\Pi T_\mu^{(\lambda)}$.

Proof) Similar to the proof of Proposition 13.

Refer to the text book.

LSTD(λ), LSPE(λ) and TD(λ)

LSTD(λ) method

$$r_k = \left(C_k^{(\lambda)}\right)^{-1} d_k^{(\lambda)}$$

LSPE(λ) method

$$r_{k+1} = r_k - \gamma G_k \left(C_k^{(\lambda)} r_k - d_k^{(\lambda)}\right), \quad k = 0, 1, 2, \cdots$$

where γ is the stepsize and G_k is a scaling matrix that converges to some G such that $I - \gamma G_k C^{(\lambda)}$ is smaller than 1.

spectral radius of

- One possibility is to choose $\gamma = 1$ and

$$G_k = \left(\frac{1}{k+1}\sum_{t=0}^{k} \phi(i_t)\phi(i_t)'\right)^{-1}$$

where one can show that $G_k \to G$ as $C_k^{(\lambda)} \to C^{(\lambda)}$ and $d_k^{(\lambda)} \to d^{(\lambda)}$ due to the law of large numbers argument.

- Introduce a vector called <u>eligibility vector</u>

$$z_t = \sum_{m=0}^{t} (\alpha\lambda)^{t-m} \phi(i_m) \qquad \text{(40)}$$

which is the weighted sum of the present and past feature vectors $\phi(i_m)$ from the simulations. Then, one can show that $C_k^{(\lambda)} \to C^{(\lambda)}$ and $d_k^{(\lambda)} \to d^{(\lambda)}$ as $k \to \infty$ where

$$C_k^{(\lambda)} = \frac{1}{k+1}\sum_{t=0}^{k} z_t \left(\phi(i_t) - \alpha\phi(i_{t+1})\right)' \qquad \text{(41)}$$

and

$$d_k^{(\lambda)} = \frac{1}{k+1}\sum_{t=0}^{k} z_t\, g(i_t, i_{t+1})$$

• In particular, (40) and (41) can be rewritten by means of recursive formulas in the form of stochastic approximation as

$$z_k = \alpha\lambda z_{k-1} + \phi(i_k)$$

$$C_k^{(\lambda)} = (1-\delta_k) C_{k-1}^{(\lambda)} + \delta_k z_k (\phi(i_k) - \alpha\phi(i_{k+1}))'$$

$$d_k^{(\lambda)} = (1-\delta_k) d_{k-1}^{(\lambda)} + \delta_k z_k g(i_k, i_{k+1})$$

with the initial conditions $z_{-1}=0$, $C_{-1}=0$, $d_{-1}=0$ and $\delta_k = \frac{1}{k+1}$, $k=0,1,\cdots$.

— (42)

• Alternatively, LSPE(λ) can be written as

$$r_{k+1} = r_k - \frac{\gamma}{k+1} G_k \sum_{t=0}^{k} z_t g_{k,t} \qquad \text{— (43)}$$

where $g_{k,t}$ is the temporal difference

$$g_{k,t} = \phi(i_t)' r_k - \alpha\phi(i_{t+1})' r_k - g(i_t, i_{t+1}).$$

TD(λ) method

- A single sample approximation of the LSPE(λ) in (43) with $G_k = I$, i.e.,

$$r_{k+1} = r_k - \frac{\gamma}{k+1} G_k \sum_{t=0}^{k} z_t g_{k,t}$$

$$\simeq r_k - \gamma_k z_k g_{k,k} \quad \text{with} \quad \gamma_k = \frac{\gamma}{k+1}.$$