Two Issues in Policy Iteration

(exploration in policy evaluation step
  oscillation in policy improvement step

Exploration

- To evaluate μ, need to generate samples using μ which biases the simulation by underrepresenting states (states with small ξ_i) that are unlikely to occur under μ.
- As a result, the cost-to-go estimates of these underrepresented states will be highly inaccurate, and seriously impact the next improved policy $\bar{\mu}$.
- The irreducibility of P_μ of μ may be hard or impossible to guarantee.
- Need to change the sampling mechanism and the simulation formulas.
- Consider a modified transition probability matrix which is irreducible

$$\bar{P}_\mu = (I - B) P_\mu + BQ$$

where B is a diagonal matrix which diagonal components $\beta_i \in [0, 1]$ and Q is another transition probability matrix.

(Why? $\bar{P}_\mu$ could address both the underrepresented state issue and the difficulty with multiple recurrent classes and transient states)

- At state i, the next state is generated with probability $1-\beta_i$ according to $p_{ij}(\mu(i))$, and with probability β_i according to q_{ij}.

↳ fictitious transition ⇒ "exploration"

c.f.) exploration vs exploitation

- Consider Projected Bellman equation for $\bar{P}_\mu$

$$\Phi r = \bar{\Pi}\, \bar{T}_\mu^{(\lambda)}\, \Phi r$$

where $\bar{T}_\mu^{(\lambda)} J = \bar{g}_\mu^{(\lambda)} + \alpha \bar{P}_\mu^{(\lambda)} J$ and $\bar{\Pi}$ denotes the projection on S w.r.t. $\|\cdot\|_{\bar{\xi}}^2$ where $\bar{\xi}$ is the steady-state distribution corresponding to $\bar{P}_\mu$.

- However, this projected Bellman equation is problematic because it solves r^* such that $\Phi r^* \sim \bar{T}_\mu^{(\lambda)} \Phi r^*$, not $\Phi r^* \sim T_\mu^{(\lambda)} \Phi r^*$.

- Instead, consider the following modified projected Bellman eq. for $\bar{P}_\mu$

$$\boxed{\Phi r = \bar{\Pi}\, T_\mu^{(\lambda)}\, \Phi r} \quad \text{---- (44)}$$

"exploration-enhanced projected Bellman equation"

where $T_\mu^{(\lambda)} J = g_\mu^{(\lambda)} + \alpha P_\mu^{(\lambda)} J$ and $\bar{\Pi}$ denotes the projection on S w.r.t. $\|\cdot\|_{\bar{\xi}}^2$ where $\bar{\xi}$ is the steady-state distribution corresponding to $\bar{P}_\mu$.

Exploration using Extra Transitions

- Apply only to $\lambda = 0$ (single-step Bellman equation) cases.
- Generate a state sequence $\{\hat{\imath}_0, \hat{\imath}_1, \cdots\}$ according to any steady-state distribution $\bar{\xi}$ (more balanced one than ξ such as the uniform distribution).
- Generate a sequence of "independent" transitions $\{(\hat{\imath}_0, j_0), (\hat{\imath}_1, j_1), \cdots\}$ according to the original transition matrix P_μ.
- The exploration-enhanced projected Bellman equation (44) yields

$$Cr = d$$

where $C = \Phi' M_{\bar{\xi}} (I - \alpha P_\mu) \Phi$, $d = \Phi' M_{\bar{\xi}} g_\mu$.

- One can derive the estimators for C and d by invoking the law of large numbers argument

$$C_k = \frac{1}{k+1}\sum_{t=0}^{k}\phi(i_t)(\phi(i_t)-\alpha\phi(j_t))'$$

$$d_k = \frac{1}{k+1}\sum_{t=0}^{k}\phi(i_t)\,g(i_t,j_t)$$

- Thus, we have

LSTD(0):

$$\sum_{t=0}^{k}\phi(i_t)\,\tilde{q}_{k,t}=0$$

where $\tilde{q}_{k,t}=\phi(i_t)'r_k-\alpha\phi(j_t)'r_k-g(i_t,j_t)$

LSPE(0):

$$r_{k+1}=r_k-\frac{\gamma}{k+1}G_k\sum_{t=0}^{k}\phi(i_t)\,\tilde{q}_{k,t}$$

TD(0):

$$r_{k+1}=r_k-\frac{\gamma}{k+1}\phi(i_k)\,\tilde{q}_{k,k}.$$

Exploration using Modified Temporal Differences

- An alternative exploration approach that works for all $\lambda\geq 0$.
- Use explicit knowledge of the transition probabilities $p_{ij}(M(i))$ and $\bar{p}_{ij}$, so it is not a model-free approach.
- Generate a single state sequence $\{i_0, i_1, \cdots\}$ according to the exploration-enhanced transition matrix $\bar{P}$ (i.e., it is a off-policy approach).

- Define a modified exploration-enhanced temporal differences

$$\bar{q}_{k,t} = \phi(i_t)' r_k - \frac{p_{i_t i_{t+1}}}{\bar{p}_{i_t i_{t+1}}} \left(\alpha \phi(i_{t+1})' r_k + g(i_t, i_{t+1}) \right) \quad \text{---- (45)}$$

where p_{ij} and $\bar{p}_{ij}$ denote the ijth components of P_u and $\bar{P}$, respectively.

- Note that the approximation of an expected value w.r.t. a given distribution (induced by the transition matrix P_u) by sampling w.r.t. a different distribution (induced by the exploration-enhanced transition matrix $\bar{P}$) is reminiscent of "importance sampling". The probability ratio $\frac{p_{i_t i_{t+1}}}{\bar{p}_{i_t i_{t+1}}}$ in (45) provides the necessary correction.

- One can derive the estimator for C and d using the modified temporal differences in a recursive form

$$C_k^{(\lambda)} = (1-\delta_k) C_{k-1}^{(\lambda)} + \delta_k z_k \left(\phi(i_k) - \alpha \frac{p_{i_k i_{k+1}}}{\bar{p}_{i_k i_{k+1}}} \phi(i_{k+1}) \right)'$$

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

where z_k are modified eligibility vectors given by

$$z_k = \alpha\lambda \frac{p_{i_{k-1} i_k}}{\bar{p}_{i_{k-1} i_k}} z_{k-1} + \phi(i_k)$$

with $z_{-1} = 0$, $C_{-1} = 0$, $d_{-1} = 0$ and $\delta_k = \frac{1}{k+1}$, $k = 0, 1, \ldots$.

Contraction Properties of Exploration-Enhanced Methods

$$\underline{\underline{\Phi}} r = \overline{\overline{\Pi}} T^{(\lambda)} \underline{\underline{\Phi}} r$$

Proposition 16.

Assume that $\overline{P}$ is irreducible and $\overline{\xi}$ is its invariant distribution. Then $T^{(\lambda)}$ and $\overline{\Pi} T^{(\lambda)}$ are contractions w.r.t. $\|\cdot\|_{\overline{\xi}}$ for all $\lambda \in [0,1)$ provided $\overline{\alpha} < 1$ where

$$\overline{\alpha} = \frac{\alpha}{\sqrt{1 - \min_{i=1,\cdots,n} \beta_i}}.$$

The associated modulus of contraction is at most equal to

$$\frac{\overline{\alpha}(1-\lambda)}{1-\overline{\alpha}\lambda}.$$

Proof) Refer to the ~~text~~ book.