NOTES ON OPTIMISTIC POLICY ITERATION WITH NONUNIFORM DISTRIBUTION

YUANLONG CHEN

In this short note, we try to extend the convergence result of optimistic policy iteration to a variant algorithm in which the initial states at each update step are selected subject to some nonuniform distribution.

1. A postive result

We use the same notations as in [1] and consider Monte Carlo method. Suppose at each iteration t, we randomly pick state $i \in \{1, \dots, n\}$ with probability $p(i) > 0$, then generate a trajectory starting with i and update the cost-to-go function $J_t(i)$ accordingly. Then we have

$$J_{t+1}(i) = \begin{cases} (1-\gamma_t)J_t(i) + \gamma_t(J^{\mu_t} + \omega_t), & \text{with probability } p(i), \\ J_t(i), & \text{otherwise.} \end{cases}$$

Follow the argument in [1], we can rewrite the udpate rule as

$$J_{t+1} = (1-\Gamma_t)J_t + \Gamma_t(J^{\mu_t} + v_t),$$

where $\Gamma_t = \gamma_t \cdot D = \gamma_t \cdot diag\{p(1), \cdots, p(n)\}$ and v_t is a random variable such that

$$E[v_t|\mathcal{F}_t] = 0, \qquad E[\|v_t\|^2|\mathcal{F}_t] \le A + B\|J_t\|^2,$$

for some constants A and B.

Suppose α is the discounted factor, and set σ as the condition number of $diag\{p(1), \cdots, p(n)\}$, i.e.

$$\sigma = \frac{\max_i p(i)}{\min_i p(i)},$$

then we have

Proposition 1.1. *If $\sigma < \frac{1}{\alpha}$, $J_t \to J^*$ almost surely.*

Sketch of the proof. As in the proof in [1], set $X_t = TJ_t - J_t$ the key is to show that

$$\limsup_{t\to\infty} X_t \le 0.$$

Define $c_t = \|TJ_t - J_t\|_\infty = \max_i |TJ_t(i) - J_t(i)|$, we can show that

$$\|J^{\mu_t} - J_t\|_\infty \le \frac{c_t}{1-\alpha} e,$$

where e is a vector with all component equal to 1. Redo the calculation as in (2) in [1], since Γ_t do not commute with P_{μ_t}, we have an extra commutator term R_t,

$$TJ_{t+1} - J_{t+1} \le (1-\Gamma_t)(TJ_t - J_t) + \Gamma_t R_t + \Gamma_t \tilde{v}_t, \tag{1.1}$$

where

$$R_t = \alpha\Gamma_t^{-1}[P_{\mu_t}, \Gamma_t](J_t - J^{\mu_t}),$$

and $\tilde{v}_t$ is a random variable such that

$$E[\tilde{v}_t|\mathcal{F}_t] = 0, \qquad E[\|\tilde{v}_t\|^2|\mathcal{F}_t] \le A + B\|J_t\|^2.$$

A straight foward calculation shows that

$$\Gamma_t^{-1}[P_{\mu_t}, \Gamma_t] = \left(p_{i,j}^{\mu_t} \left(\frac{p(j)}{p(i)} - 1 \right) \right),$$

note that

$$\left| \frac{p(j)}{p(i)} - 1 \right| \leq \sigma - 1 < \frac{1}{\alpha} - 1 = \frac{1-\alpha}{\alpha},$$

we then see that

$$\begin{aligned} |R_t(i)| &= \left| \alpha \sum_{j=1}^{n} p_{i,j}^{\mu_t} \left(\frac{p(j)}{p(i)} - 1 \right) (J_t(j) - J^{\mu_t}(j)) \right| \\ &\leq \sum_{j=1}^{n} p_{i,j}^{\mu_t} |\sigma - 1| \frac{\alpha}{1-\alpha} c_t \\ &= \beta c_t. \end{aligned}$$

where $\beta = \frac{|\sigma-1|\cdot\alpha}{1-\alpha} < 1$. Together with(1.1), this implies

$$X_{t+1} \leq (1 - \Gamma_t) X_t + \Gamma_t (\beta \|X_t\|_\infty \cdot e) + \Gamma_t \tilde{v}_t.$$

Using the same argument in [1], we can see that

$$\limsup_{t \to \infty} X_t \leq 0.$$

The rest of proof is identical to [1]. □

2. Optimality

Question: is the condition $\sigma < \frac{1}{\alpha}$ sharp?

I am thinking about to investigate the dynamics of $TJ_t - J_t$ from the perspective of ODE approach. Set $X_t = TJ_t - J_t$, i.e.

$$g_{\mu_t} + \alpha P_{\mu_t} J_t - J_t = X_t.$$

Since $\alpha P_{\mu_t} - I$ is invertible (its eigenvalues are no-zeros), solving this matrix equation, we have

$$J_t = (\alpha P_{\mu_t} - I)^{-1} X_t - (\alpha P_{\mu_t} - I)^{-1} g_{\mu_t}. \tag{2.1}$$

On the other hand, note J^{μ_t} is the fixed point of T_{μ_t}, i.e.

$$g_{\mu_t} + \alpha P_{\mu_t} J^{\mu_t} = J^{\mu_t},$$

we have

$$J^{\mu_t} = -(\alpha P_{\mu_t} - I)^{-1} g_{\mu_t}. \tag{2.2}$$

From (2.1) and (2.2), we see that

$$J_t - J^{\mu_t} = (\alpha P_{\mu_t} - I)^{-1} X_t$$

Equation (1.1) then, after some simplification, becomes

$$X_{t+1} \leq X_t + \gamma_t (H_t X_t + \tilde{v}_t), \tag{2.3}$$

where

$$H_t = \alpha [P_{\mu_t}, D] (\alpha P_{\mu_t} - I)^{-1} - D.$$

If H_t is negative definte no matter what policy μ_t is, the convergence result is easy to establish, but this might be true only under certain conditions(such as $\sigma < 1/\alpha$). For

general non-uniform distribution, it's too much to hope this is still true. Intuitively, H_t should satisfy some other weaker properties, e.g., its eigenvalues have negative real part, however, when distribution is non-uniform, the trajactory of X_t is curved and may cause the oscillation of policies. The convergence results are not easy to establish from this point of view. But can we construct any counterexample to show that $\sigma < 1/\alpha$ is sharp? I am still thinking about it...

References

1. John N. Tsitsiklis, *On the convergence of optimistic policy iteration*, J. Mach. Learn. Res. **3** (2003), no. 1, 59–72. MR 1966053

University of Washington, Seattle
Email address: ylchen88@uw.edu