Whittle Index-based Q-learning using FGDQN

Tejas Pagare

February 2, 2022

Algorithm

Consider parametrized families $\lambda(k;\sigma)$ and $Q(i,u,\lambda;\theta)$ where we render the implicit dependence of Q on λ by taking it as a input to the Q-network.

For Q iterate, we consider a single run of a simulated controlled Markov chain $(X_n, U_n), n \geq 0$ so that X_{n+1} have the conditional law as $p(\cdot|X_n, U_n)$.

$$\begin{aligned}
\theta_{n+1} \;=\; \theta_n \;-\; a(n)\Big(&\nabla_\theta Q(X_{n+1}, v_n, \lambda(X_{n+1};\sigma_n);\theta_n) \;-\; \nabla_\theta f(Q(X_n, U_n, \lambda(X_n;\sigma_n);\theta_n)) \\
&- \nabla_\theta Q(X_n, U_n, \lambda(X_n;\sigma_n);\theta_n)\;\Big) \times \\
&\overline{\Big((1-U_n)(r(X_n,0)+\lambda(X_n;\sigma_n)) + U_n r_n(X_n,1) + \max_{v\in\{0,1\}} Q(X_{n+1}, v, \lambda(X_{n+1};\sigma_n);\theta_n)} \\
&\overline{-\; f(Q(X_n, U_n, \lambda(X_n;\sigma_n);\theta_n)) - Q(X_n, U_n, \lambda(X_n;\sigma_n);\theta_n)}\Big) + a(n)\xi_{n+1},
\end{aligned} \tag{1}$$

The term with the overline comprises of averaging at time n over past traces sampled from $(X_k, U_k, X_{k+1}), k \leq n$, for which, $X_k = X_n$ & $U_k = U_n$.

For whittle iterate, we consider a single run of a simulated controlled Markov chain $(X_n, 0), n \geq 0$ so that X_{n+1} have the conditional law as $p(\cdot|X_n, 0)$, as required in the derivation.

$$\begin{aligned}
\sigma_{n+1} = \sigma_n - b(n)&\overline{\Big(Q(X_n, 1, \lambda(X_n;\sigma_n);\theta_n) - r(X_n,0) + \; f(Q(X_n, 0, \lambda(X_n;\sigma_n);\theta_n))} \\
&\overline{- \max_{v\in\{0,1\}} Q(X_{n+1}, v, \lambda(X_{n+1};\sigma_n);\theta_n) - \lambda(X_n;\sigma_n)\Big)} \times \\
&\Big(\nabla_\lambda Q(X_n, 1, \lambda(X_n);\theta_n)\nabla_\sigma \lambda(X_n;\sigma_n) \;+\; \nabla_\sigma f(Q(X_n, 0, \lambda(X_n;\sigma_n);\theta_n)) \\
&- \nabla_\sigma Q(X_{n+1}, v_n, \lambda(X_n;\sigma_n);\theta_n) - \nabla_\sigma \lambda(X_n;\sigma_n)\Big)
\end{aligned} \tag{2}$$

The term with the overline comprises of averaging at time n over past traces sampled from $(X_k, U_k, X_{k+1}), k \leq n$, for which, $X_k = X_n$ & $U_k = 0$.

Please check next page for the derivation $\longrightarrow$

Derivation

As we have seen, whittle index is equivalent to solving for $\lambda(X_n)$ the following equation

$$Q(X_n, 1) = Q(X_n, 0)$$

Substituting $Q(X_n, 0)$ we get

$$Q(X_n, 1) = r(X_n, 0) + \lambda(X_n) - \rho + \sum p(X_{n+1}|X_n, 0) \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n)$$

which is equivalent to

$$\lambda(X_n) = Q(X_n, 1) - r(X_n, 0) + \rho - \sum p(X_{n+1}|X_n, 0) \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n)$$

Now, using stochastic approximation we remove the conditional expectation by a real random variable ξ_{i0} with the law $p(\cdot|X_n, 0)$ and make increment based on our current estimate.
Hence, we consider a single run $\{X = X_n, U = 0\}$ of the controlled Markov chain which gives X_{n+1} with the same conditional law of $p(\cdot|X_n, 0)$.
$\therefore$ We now know have,

$$\lambda(X_n) = (1 - b(n))\lambda(X_n) + \Big(Q(X_n, 1) - r(X_n, 0) + \rho - \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n)\Big)$$

We replace ρ with it's current estimate i.e. $f(Q_n)$ to get

$$\lambda(X_n) = (1 - b(n))\lambda(X_n) + \Big(Q(X_n, 1) - r(X_n, 0) + f(Q_n) - \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n)\Big)$$

$$\begin{aligned} \lambda(X_n) = \lambda(X_n) + b(n)\Big(&Q(X_n, 1) - r(X_n, 0) \\ &+ f(Q_n) - \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n) - \lambda(X_n))\Big) \end{aligned} \tag{3}$$

Following we get the iteration for Whittle Index parameters σ

$$\begin{aligned} \sigma_{n+1} = \sigma_n + b(n) \times \Big(&Q(X_n, 1; \theta_n) - r(X_n, 0) \\ &+ f(Q_n) - \max_{v \in \{0,1\}} Q(X_{n+1}, v; \theta_n) - \lambda(X_n; \sigma_n)\Big)\nabla_\sigma \lambda(X_n; \sigma_n) \end{aligned} \tag{4}$$

Similar derivation can be done when we consider implicit dependence of Q on λ by taking it as one of the inputs.