LINMA2725 Stochastic Optimal Control and Reinforcement Learning Part III

Course 3: Actor-Critic Methods

Guillaume Berger

November 25, 2024

Reference: [1], Chapter 10.

Any questions or feedback are welcome.

Stochastic Optimal Control and Reinforcement Learning

Part III: Stochastic Systems

Guillaume Berger

Table of content

- Stochastic systems and stochastic control (1 course)
- Learning techniques for stochastic control (1-2 courses)
  - Variance reduction techniques
  - Actor-critic methods
- Online learning techniques for stochastic control (1-2 courses)

Problem setting and reminders

Consider a given (randomized) policy $\breve{\phi}(u|x)$

Let ϖ be the steady-state distribution of the closed-loop system

Let $Q(x,u)$ be its Q-function and $h(x)$ its value function

Cost can be total, discounted or average cost

Objective: approximate Q or similar quantities

Assume: $Q^{\theta}(x,u) = \theta^{\mathsf{T}}\psi(x,u)$ (linear template)

For theoretical analysis: assume X and U finite

Advantage function

Advantage function: $V(x,u) = Q(x,u) - h(x)$

Motivation: $h = \arg\min_{g:X\to\mathbb{R}} \mathbb{E}_{\varpi}\left[\left\{Q\big(\Phi(n)\big) - g\big(X(n)\big)\right\}^2\right]$

Hence, L_2-norm of V is smaller than Q

But optimal policy of V and Q are the same!

Note that $\mathbb{E}_{\varpi}\left[\, V\big(\Phi(n)\big) \mid X(n) \,\right] = 0$

The idea is that if we "offset" $Q(x,u)$ with a function $g(x)$, then this does not change the associated optimal policy, i.e.,

$$\arg\min_{u\in\mathsf{U}} Q(x,u) = \arg\min_{u\in\mathsf{U}} \{Q(x,u) - g(x)\}.$$

We choose the offset $g(x)$ that minimizes the L^2-norm of $Q(x,u)-g(x)$, which is a sensible idea since having quantities with a smaller magnitude (norm) involved in the recursive algorithm will generally induce a smaller variance of the algorithm. The optimal $g(x)$ is given by $h(x) \triangleq \sum_{u\in\mathsf{U}} \breve{\phi}(u|x)$.

Advantage TD(λ)-learning

TD(λ) algorithm (on-policy for advantage)

For initialization $\omega_0\,, \zeta_0 \in \mathbb{R}^m$, the sequence of estimates are defined recursively:

$$
\begin{aligned}
\omega_{n+1} &= \omega_n + \alpha_{n+1}\zeta_n \mathcal{D}_{n+1} \\
\mathcal{D}_{n+1} &= \big(-H^{\omega}(\Phi(n)) + c_n + \gamma H^{\omega}(\Phi(n+1))\big)\Big|_{\omega=\omega_n} \\
\zeta_{n+1} &= \lambda\gamma\zeta_n + \widetilde{\psi}^{\circ}(\Phi(n+1))\,, \qquad H^{\omega}(\Phi(n)) = \omega^{\intercal}\widetilde{\psi}^{\circ}(\Phi(n))
\end{aligned}
\tag{9.68}
$$

where $\underline{\psi}(x) = \sum_u \psi(x,u)\breve{\phi}(u \mid x)$ *and* $\widetilde{\psi}(x,u) = \psi(x,u) - \underline{\psi}(x)\,,$

Same as TD(λ)-learning but with a special basis!

Motivation of advantage TD(λ)-learning

Remember $\lambda = 1$: θ^* *solves*

$$\theta^* = \arg\min_\theta \|H^\theta - Q\|^2_\varpi \stackrel{\text{def}}{=} \sum_{x \in \mathsf{X},\, u \in \mathsf{U}} (H^\theta(x,u) - Q(x,u))^2 \varpi(x,u)$$

New result: the projection of V on $\left\{\theta^{\mathsf{T}}\psi + \xi^{\mathsf{T}}\underline{\psi}\right\}$ is equal to the projection of Q on $\left\{\theta^{\mathsf{T}}\tilde{\psi}\right\}$

Hence, H^{θ^*} is the projection of V on $\left\{\theta^{\mathsf{T}}\psi + \xi^{\mathsf{T}}\underline{\psi}\right\}$

Proof of the result. Let $\hat{Q}$ be the projection of Q on $\{\theta^\top \tilde{\psi} : \theta \in \mathbb{R}^d\}$. We show that $V - \hat{Q}$ is orthogonal to ψ and $\underline{\psi}$. This will imply that $\hat{Q}$ is the projection of V on $\{\theta^\top \psi + \xi^\top \underline{\psi} : \theta, \xi \in \mathbb{R}^d\}$. We show the orthogonality only for ψ since the proof for $\underline{\psi}$ is easier. For that, observe that

$$\begin{aligned}
\mathsf{E}_\varpi \left[(V - \hat{Q})\psi\right] &= \mathsf{E}_\varpi \left[(Q - h - \hat{Q})(\tilde{\psi} + \underline{\psi})\right] \\
&= \mathsf{E}_\varpi \left[(Q - \hat{Q})\tilde{\psi}\right] + \mathsf{E}_\varpi \left[Q\underline{\psi}\right] - \mathsf{E}_\varpi \left[\hat{Q}\underline{\psi}\right] - \mathsf{E}_\varpi \left[h\tilde{\psi}\right] - \mathsf{E}_\varpi \left[h\underline{\psi}\right] \\
&\quad \text{the first term is zero by definition of } \hat{Q}, \\
&\quad \text{the third and fourth terms are both zero because } \mathsf{E}_\varpi \left[\tilde{\psi}(\Phi(n)) \,\middle|\, X(n)\right] = 0 \\
&\quad \text{the second is equal to } \mathsf{E}_\varpi \left[h\underline{\psi}\right] \text{ since } \mathsf{E}_\varpi \left[Q(\Phi(n)) \,|\, X(n)\right] = h(X(n)) \\
&= \mathsf{E}_\varpi \left[h\underline{\psi}\right] - \mathsf{E}_\varpi \left[h\underline{\psi}\right] = 0.
\end{aligned}$$

This concludes the proof. □

Advantage TD(λ)-learning as a LS system

LSTD(1) for advantage with weighting (on-policy)

With weighting function $w\colon \mathsf{X} \to (0,\infty)$, initialization $\zeta_0 \in \mathbb{R}^m$, $\widehat{\Sigma}_0 \in \mathbb{R}^{m\times m}$, and time horizon N,

$$\omega_N = \widehat{\Sigma}_N^{-1}\bar{\psi}_N^Q \tag{10.16a}$$

with

$$\widehat{\Sigma}_N = \frac{1}{N}\Big(\widehat{\Sigma}_0 + \sum_{n=1}^{N} w_n \widetilde{\psi}_{(n)}\widetilde{\psi}_{(n)}^{\intercal}\Big) \tag{10.16b}$$

$$\bar{\psi}_N^Q = \frac{1}{N}\sum_{n=1}^{N} c_n\zeta_n \tag{10.16c}$$

$$\zeta_n = \gamma\zeta_{n-1} + w_n\widetilde{\psi}_{(n)}\,, \quad w_n = w(X(n))\,, \quad \widetilde{\psi}_{(n)} = \widetilde{\psi}^{\circ}(\Phi(n))\,, \qquad 1 \le n \le N \tag{10.16d}$$

Regeneration

Let $z^\bullet = (x^\bullet, u^\bullet) \in \mathsf{X} \times \mathsf{U}$ denote any state with positive steady-state probability: $\varpi(z^\bullet) > 0$.

$$\widetilde{Q}_\gamma(z) = \mathsf{E}\Big[\sum_{k=0}^{\tau_\bullet - 1} \gamma^k \tilde{c}(\Phi(k)) \mid \Phi(0) = z\Big] + \mathsf{E}\Big[\gamma^{\tau_\bullet} \sum_{k=0}^{\infty} \gamma^k \tilde{c}(\Phi(k + \tau_\bullet)) \mid \Phi(0) = z\Big]$$

$\tau_\bullet = \min\{ k \geq 1 \mid \Phi(k) = z^\bullet \}$

Result: $\widetilde{Q}_\gamma(z) = \mathsf{E}\Big[\sum_{k=0}^{\tau_\bullet - 1} \gamma^k \tilde{c}(\Phi(k)) \mid \Phi(0) = z\Big] + \widetilde{Q}_\gamma(z^\bullet)\mathsf{E}\big[\gamma^{\tau_\bullet} \mid \Phi(0) = z\big]$

focus on this to reduce variance!

tends to constant when $\gamma \to 1$

Focusing on the first term (orange) is another way of removing an offset without impacting the associated optimal policy because the second term (blue) is assumed to be constant (independent of z).

When $\gamma = 1$, regeneration is also crucial to have boundedness of the eligibility vectors ζ_n in the TD(1)-learning algorithm, as we will see next.

Regeneration for average cost

Define $H_3(z) = \mathsf{E}\Big[\sum_{k=0}^{\tau_\bullet - 1} \tilde{c}(\Phi(k)) \mid \Phi(0) = z\Big]$

Goal: approximate H_3 with H^θ

Motivations:

- Average cost is important for stochastic systems!
- Other costs like mean discounted cost can be formulated as average cost

The average cost can be used as a metric to compare policies between them because it is a scalar. By contrast, the discounted cost $h_\gamma(x)$ cannot be used as a metric because it is a function. Alternatively, one can use the "mean discounted cost": given a probability distribution μ on X, define the metric $\langle \mu, h_\gamma \rangle \triangleq \sum_{x \in \mathsf{X}} \mu(x) h_\gamma(x)$.

The mean discounted cost of a given system can be formulated as the average cost of another system constructed from the original one; see [1, § 10.4.1] for such a construction.

Regenerative TD(λ)-learning for average cost

Let $z^\bullet = (x^\bullet, u^\bullet) \in \mathsf{X} \times \mathsf{U} : \varpi(z^\bullet) > 0.$

Regenerative TD(λ) algorithm for average cost (on-policy)

For initialization $\theta_0\,, \zeta_0 \in \mathbb{R}^d$, the sequence of estimates are defined recursively:

$$
\begin{aligned}
\theta_{n+1} &= \theta_n + \alpha_{n+1}\zeta_n \mathcal{D}_{n+1} \\
\mathcal{D}_{n+1} &= \big(-H^\theta(\Phi(n)) + \tilde{c}_n + \mathbb{1}\{\Phi(n+1) \neq z^\bullet\} H^\theta(\Phi(n+1))\big)\Big|_{\theta=\theta_n} \\
\zeta_{n+1} &= \lambda \mathbb{1}\{\Phi(n+1) \neq z^\bullet\}\zeta_n + \psi_{(n+1)} \\
\eta_{n+1} &= \eta_n + \tilde{c}_n/(n+1)\,, \qquad \tilde{c}_n = c(\Phi(n)) - \eta_n \qquad\qquad n \geq 0
\end{aligned}
\tag{10.28}
$$

Analysis of regenerative TD(λ)-learning

This is a linear SA algorithm

$$A_{n+1} = \zeta_n \big[-\psi_{(n)} + \mathbb{1}\{\Phi(n+1) \neq z^\bullet\} \psi_{(n+1)} \big]^\intercal$$
$$b_{n+1} = -\zeta_n [c_n - \eta_n]$$

The associated ODE is thus linear, with vector field

$$\bar{f}(\vartheta) = A\vartheta + b\,, \qquad A = \mathsf{E}_\varpi[A_n]\,, \quad b = -\mathsf{E}_\varpi[\zeta_n \tilde{c}(\Phi(n))]$$

Theorem 10.14. (L_2 Optimality of TD(1) for average cost) *Consider the algorithm (10.28) with linear function approximation $H^\theta = \theta^\intercal \psi$. Assume that $\mathbf{\Phi}$ is uni-chain and that $\varpi(z^\bullet) > 0$.*

Then, in the special case $\lambda = 1$,

(i) $A = -R(0)$

(ii) *Any solution to $0 = \bar{f}(\theta^*) = \mathsf{E}_\varpi[\zeta_n \mathcal{D}_{n+1}]$ solves the minimum norm problem:*

$$\theta^* \in \arg\min_\theta \|H^\theta - H_3\|_\varpi^2 = \arg\min_\theta \mathsf{E}_\varpi\big[\big(H^\theta(\Phi(n)) - H_3(\Phi(n))\big)^2\big] \tag{10.29}$$

See [1, Theorem 10.14]. Reminder: $R(0) \triangleq \mathsf{E}_{\varpi}\left[\psi(\Phi(n))\psi(\Phi(n))^{\top}\right]$ is positive definite (under mild assumption of linear independence of ψ) so that A is Hurwitz.

Regenerative relative TD(λ)-learning

Regenerative relative TD(λ) algorithm for average cost (on-policy)

For initialization $\theta_0\,, \zeta_0 \in \mathbb{R}^d$, the sequence of estimates are defined recursively:

$$\theta_{n+1} = \theta_n + \alpha_{n+1}\zeta_n \mathcal{D}_{n+1}$$

$$\mathcal{D}_{n+1} = \big(-H^\theta(\Phi(n)) - \delta\langle \mu\,, H^\theta\rangle + c_n + H^\theta(\Phi(n+1))\big)\Big|_{\theta=\theta_n} \tag{10.32}$$

$$\zeta_{n+1} = \lambda \mathbb{1}\{\Phi(n+1) \neq z^\bullet\}\zeta_n + \psi_{(n+1)}$$

Variance expected to be reduced compared to non-relative version

ODE approximation $\bar{f}(\theta) = A\theta + b$ can be shown to be invertible if δ is small enough

Conditions to ensure that A is Hurwitz not available; hence, better to use LSTD(λ)

See [1, Theorem 10.15] and the discussion below it.

Actors and critics

Parametrized (randomized) policy: $\breve{\phi}^{\theta}(u|x)$

Actor

Denote

$$c_\theta(x) = \sum_u \breve{\phi}^{\theta}(u \mid x) c(x,u)$$

$$P_\theta(x,x') = \sum_u \breve{\phi}^{\theta}(u \mid x) P_u(x,x')\,,$$

$$T_\theta(z,z') = P_u(x,x')\breve{\phi}^{\theta}(u' \mid x')$$

Critic

Actor-Critic Objective.

$$\Gamma(\theta) = \sum_{x \in \mathsf{X}} c_\theta(x)\, \pi_\theta(x) = \sum_{z \in \mathsf{Z}} c(z) \varpi_\theta(z)$$

= average cost

Gradient and score function

Score function: $\Lambda^\theta(x,u) = \nabla_\theta \log[\breve{\phi}^\theta(u \mid x)]$

Note that

$$\nabla_\theta c_\theta(x) = \sum_u \breve{\phi}^\theta(u \mid x)\Lambda^\theta(x,u)c(x,u)$$

$$\nabla_\theta P_\theta(x,x') = \sum_u \breve{\phi}^\theta(u \mid x)\Lambda^\theta(x,u)P_u(x,x')\,, \qquad x,x' \in \mathsf{X},\ \theta \in \mathbb{R}^d$$

$$\Lambda^\theta(z') = \nabla_\theta \log(T_\theta(z,z'))\,, \qquad z,z' \in \mathsf{Z}$$

Gradient of critic and sensitivity theorem

Denote $Q_\theta(x,u)$ = Q-function of system controlled with $\check{\phi}^\theta$

It holds that $$\nabla\Gamma(\theta) = \mathsf{E}_{\varpi_\theta}\left[\Lambda^\theta(\Phi(k))Q_\theta(\Phi(k))\right]$$ (sensitivity theorem)

Proof: use Poisson equation $$T_\theta Q_\theta = Q_\theta - c + \Gamma(\theta).$$

Proof. Expand the Poisson equation:

$$\sum_{z' \in \mathsf{X} \times \mathsf{U}} T_\theta(z, z') Q_\theta(z') = Q_\theta(z) - c(z) + \Gamma(\theta).$$

Take gradient on both sides:

$$\sum_{z' \in \mathsf{X} \times \mathsf{U}} \nabla_\theta T_\theta(z, z') Q_\theta(z') + T_\theta(z, z') \nabla_\theta Q_\theta(z') = \nabla_\theta Q_\theta(z) + \nabla_\theta \Gamma(\theta)$$

or

$$\sum_{z' \in \mathsf{X} \times \mathsf{U}} T_\theta(z, z') \frac{\nabla_\theta T_\theta(z, z')}{T_\theta(z, z')} Q_\theta(z') + T_\theta(z, z') \nabla_\theta Q_\theta(z') = \nabla_\theta Q_\theta(z) + \nabla_\theta \Gamma(\theta)$$

Use the score function:

$$\sum_{z' \in \mathsf{X} \times \mathsf{U}} T_\theta(z, z') \Lambda^\theta(z') Q_\theta(z') + T_\theta(z, z') \nabla_\theta Q_\theta(z') = \nabla_\theta Q_\theta(z) + \nabla_\theta \Gamma(\theta)$$

Take the expectation with respect to ϖ_θ on both sides:

$$\sum_{z \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) \sum_{z' \in \mathsf{X} \times \mathsf{U}} T_\theta(z, z') \Lambda^\theta(z') Q_\theta(z') + T_\theta(z, z') \nabla_\theta Q_\theta(z') = \sum_{z \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) \nabla_\theta Q_\theta(z) + \nabla_\theta \Gamma(\theta)$$

Use that $\sum_{z \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) T_\theta(z, z') = \varpi^\theta(z')$:

$$\sum_{z' \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) \Lambda^\theta(z') Q_\theta(z') + \varpi_\theta(z') \nabla_\theta Q_\theta(z') = \sum_{z \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) \nabla_\theta Q_\theta(z) + \nabla_\theta \Gamma(\theta)$$

Change z' into z:

$$\sum_{z \in \mathsf{X} \times \mathsf{U}} \varpi^\theta(z) \Lambda^\theta(z) Q_\theta(z) = \nabla_\theta \Gamma(\theta).$$

This concludes the proof. □

Actor-critic method (ideal case)

Stochastic gradient method:

$$\theta_{n+1} = \theta_n - \alpha_{n+1} \breve{\nabla}_\Gamma(n)\,, \qquad \breve{\nabla}_\Gamma(n) \stackrel{\text{def}}{=} \Lambda^{\theta_n}(\Phi(n)) Q_{\theta_n}(\Phi(n))$$

Challenges:

- Q_θ is not known
- Variance can be large

... addressed next

Actor-critic method

Idea: $H_\theta^\omega \approx Q_\theta$ where $H_\theta^\omega = \omega^\intercal \psi_\theta\,, \qquad \omega \in \mathbb{R}^{d'}\,, \quad \theta \in \mathbb{R}^d$

Actor-Critic Algorithm

For initialization $\theta_0 \in \mathbb{R}^d$ and $\omega_0\,, \zeta_0 \in \mathbb{R}^{d'}$,

$$\theta_{n+1} = \theta_n - \alpha_{n+1}\breve{\nabla}_\Gamma(n)\,, \qquad \breve{\nabla}_\Gamma(n) \stackrel{\text{def}}{=} \Lambda^{\theta_n}(\Phi(n))H_{\theta_n}^{\omega_n}(\Phi(n)) \tag{10.44a}$$

$$\Phi(n+1) \sim T_{\theta_n}(z, \cdot\,)\,, \quad \textit{with } z = \Phi(n) \tag{10.44b}$$

$$\left.\begin{aligned}
\mathcal{D}_{n+1} &= \{-H_\theta^\omega(\Phi(n)) + \tilde{c}_n + \mathbb{1}\{\Phi(n+1) \neq z^\bullet\}H_\theta^\omega(\Phi(n+1))\}\Big|_{\substack{\theta=\theta_n \\ \omega=\omega_n}} \\
\omega_{n+1} &= \omega_n + \beta_{n+1}\zeta_n\mathcal{D}_{n+1} \\
\zeta_{n+1} &= \lambda\mathbb{1}\{\Phi(n+1) \neq z^\bullet\}\zeta_n + \psi_{\theta_{n+1}}(\Phi(n+1)) \\
\eta_{n+1} &= \eta_n + \beta_{n+1}\tilde{c}_n\,, \qquad \tilde{c}_n = c(\Phi(n)) - \eta_n
\end{aligned}\right\} \tag{10.44c}$$

Regenerative TD(λ)

Analysis of actor-critic method

Proposition 10.17. *Suppose that* $\lambda = 1$ *and the step-size sequences satisfy* $\lim_{n\to\infty} \frac{\beta_n}{\alpha_n} = \infty$

Assume the following consistency condition holds:

$$\text{for each } \theta \in \mathbb{R}^n \text{ there is a } \omega_\theta^* \in \mathbb{R}^{d'} \text{ satisfying } H_\theta^{\omega_\theta^*} = Q_\theta. \tag{10.45}$$

Then, the ODE approximation of (10.44) *is gradient descent* $\frac{d}{dt}\vartheta = -\nabla\Gamma(\vartheta)$. □

Consistency assumption (10.45) can be restrictive; remedy on next slide

See [1, Proposition 10.17].

Relaxing the consistency assumption

Assume: Compatible Features Property (CFP)

$$\Lambda_i^\theta \in \mathcal{H}^\theta \qquad for\ each\ \theta \in \mathbb{R}^d\ and\ 1 \le i \le d$$

(can always be done by design of $\mathcal{H}^\theta$)

Proposition 10.18. *let* $\widehat{Q} \in \arg\min\{\|H - Q_\theta\|^2_{\varpi_\theta} : H \in \mathcal{H}^\theta\}$

Then,

$$\nabla\Gamma(\theta) = \mathsf{E}_{\varpi_\theta}\left[\Lambda^\theta(\Phi(k))Q_\theta(\Phi(k))\right] = \mathsf{E}_{\varpi_\theta}\left[\Lambda^\theta(\Phi(k))\widehat{Q}(\Phi(k))\right]$$

Proof. Since $\hat{Q}$ is the projection of Q_θ on $\mathcal{H}^\theta$, $\hat{Q} - Q_\theta$ is orthogonal to any function $H \in \mathcal{H}^\theta$, i.e.,

$$\mathsf{E}_{\varpi_\theta}\left[(\hat{Q} - Q_\theta)H\right] = 0.$$

By the CFP $\Lambda_i^\theta \in \mathcal{H}^\theta$, concluding the proof. □

Relaxing the consistency assumption

(continued)

Proposition 10.19. *Suppose that the assumptions of Prop. 10.17 hold, but with (10.45) replaced with the compatible features assumption (10.47).*

Then, the ODE approximation of (10.44) is unchanged: $\frac{d}{dt}\vartheta = -\nabla\Gamma(\vartheta)$. □

Hence, same conclusion but with realistic assumption (CFP)

See [1, Proposition 10.19].

Variance reduction through advantage

Result: $\textit{For any function } g\colon \mathsf{X} \to \mathbb{R}, \quad 0 = \mathsf{E}[g(X(k))\Lambda^{\theta}(\Phi(k))]$

Hence, we can subtract $G_n(X(n))$ from $H_{\theta_n}^{\omega_n}\big(\Phi(n)\big)$ without impact

For instance, use $V_\theta^\omega(x,u) \stackrel{\text{def}}{=} H_\theta^\omega(x,u) - \underline{H}_\theta^\omega(x) = \omega^\intercal \widetilde{\psi}_\theta(x,u)$

This gives: $\theta_{n+1} = \theta_n - \alpha_{n+1}\breve{\nabla}_\Gamma^v(n)\,, \qquad \breve{\nabla}_\Gamma^v(n) \stackrel{\text{def}}{=} \Lambda^{\theta_n}(\Phi(n))\underline{V_{\theta_n}^{\omega_n}(\Phi(n))}$

The idea is the same as for the TD(λ)-learning algorithm applied to the advantage function: offset the Q-function $Q_\theta(x,u)$ by some function $g_\theta(x)$ because this will not change the value of $\mathsf{E}_{\varpi_\theta}\left[\Lambda^\theta(\Phi(n))Q_\theta(\Phi(n))\right]$ but is expected to reduce the variance of the approximated gradient.

Proof of the result. Observe that

$$\mathsf{E}_{\varpi_\theta}\left[g(X(n))\Lambda^\theta(\Phi(n))\right] = \sum_{x\in\mathsf{X}} \pi_\theta(x)g(x) \sum_{u\in\mathsf{U}} \breve{\phi}^\theta(u|x)\Lambda^\theta(x,u).$$

It holds that

$$\sum_{u\in\mathsf{U}} \breve{\phi}^\theta(u|x)\Lambda^\theta(x,u) = \sum_{u\in\mathsf{U}} \nabla_\theta \breve{\phi}^\theta(u|x) = \nabla_\theta \sum_{u\in\mathsf{U}} \breve{\phi}^\theta(u|x) = \nabla_\theta(1) = 0,$$

concluding the proof. □

Variance reduction through advantage

(continued)

Using the advantage instead of the Q-function is expected to induce a smaller variance of the method

Even better: if computable, use

$$\underline{\Lambda H}_\theta^\omega(x) = \mathsf{E}[\Lambda^\theta(\Phi(n))H_\theta^\omega(\Phi(n)) \mid \mathcal{F}_n^- \,;\, X(n) = x]$$

The latter has smaller asymptotic variance than $\Lambda^{\theta_n}(\Phi(n))V_{\theta_n}^{\omega_n}(\Phi(n))$

Caveats: variance of actor-critic methods

The theoretical results were obtained for $\lambda = 1$

However, despite the proposed remedies, the variance of the actor-critic method can remain large for $\lambda = 1$

In practice, sometimes better to use $\lambda < 1$ to tame the variance

Especially for continuous systems, $\lambda = 1$ may be unapplicable

Remember that

$$\zeta_n = \sum_{k=0}^{n} \lambda^k \mathbf{1}\left\{ \bigwedge_{0 \le i \le k} \Phi(n-i) \neq z^\bullet \right\} \psi(\Phi(n-k)).$$

If $z^\bullet$ has a small steady-state probability (e.g., for continuous systems, this probability is zero in general), and $\lambda = 1$, then ζ_n can grow unbounded as $n \to \infty$. This will induce large variance in the method. Hence, even though the theoretical results (in terms of quality of the approximation of the limit point) were mostly obtained for $\lambda = 1$, in practice, it may be better to use $\lambda < 1$ to have a smaller variance.

Next course

- Online learning techniques for stochastic control
  - Bandit problem
  - Regret minimization
  - Exploration vs. exploitation trade-off

References

[1] Sean Meyn. *Control systems and reinforcement learning*. Cambridge University Press, 2022.