C8.7 Optimal Control

Sheet 2 — HT25

Section A

1. Consider the control of a process X which takes values in the discrete set $\{(i,j); i,j \in \{0,1,...,K-1\}\}$. At each time, the controller can choose either to do nothing, in which case X will take a step in a randomly chosen cardinal direction (in particular, to one of the four points of the form $(i \pm 1, j \pm 1) (\mathrm{mod}\, K)$, so this is a discrete walk on the torus). Alternatively, the controller can choose one direction, and modify the probability of walking in that direction to be a given constant $\alpha < 1$, while all other directions are chosen with equal probability.

   If the controller chooses to intervene, they must pay a cost 1 for doing so. In addition, at each time in state (i,j), they face a random state-dependent cost taken from a Binomial $(N, \frac{i+1}{i+j+1})$ distribution.

   Implement policy iteration to solve this problem.

   Solution: See attached file `PS2_policy.ipynb`

2. Prove that the SARSA iteration, as described in the lecture notes, converges.

   Solution: We follow the proof of convergence of Q-learning in the lecture notes. We define Take Q_n^π to be the sequence generated by SARSA, and Q^π to be the true Q^π-function. Define

$$\begin{aligned} Y_n(x,u) &= Q_n^\pi(x,u) - Q^\pi(x,u), \\ \alpha_n(x,u) &= \alpha_n 1_{\{(X_n,U_n)=(x,u)\}}, \\ Z_{n+1}(x,u) &= \Big[G_n + e^{-\rho} Q_n^\pi(X_{n+}, U_{n+}) - Q^\pi(X_n,u)\Big] 1_{\{(X_n,U_n)=(x,u)\}}. \end{aligned}$$

   With this notation, simple rearrangement shows that $Y_n(x,u)$ satisfies the dynamics

$$Y_{n+1}(x,u) = \big(1-\alpha_n(x,u)\big) Y_n(x,u) + \alpha_n(x,u) Z_{n+1}(x,u).$$

   From the definition of Q^π, we know that

$$0 = \mathbb{E}^u\Big[G_n + e^{-\rho} Q^\pi(X_{n+}, U_{n+}) - Q^\pi(x,u)\Big|\mathcal{F}_n\Big] 1_{\{(X_n,U_n)=(x,u)\}}$$

Sam Cohen: cohens@maths.ox.ac.uk

and thus

$$\begin{aligned}\mathbb{E}[Z_{n+1}(x,u)|\mathcal{F}_n] &= \mathbb{E}\Big[G_n + e^{-\rho}Q_n^\pi(X_{n+},U_{n+}) - Q^\pi(X_n,u)\Big|\mathcal{F}_n\Big]1_{\{(X_n,U_n)=(x,u)\}}\\ &= e^{-\rho}\mathbb{E}\Big[Q_n^\pi(X_{n+},U_{n+}) - Q(X_{n+},U_{n+})\Big|\mathcal{F}_n\Big]1_{\{(X_n,U_n)=(x,u)\}},\end{aligned}$$

Taking an absolute value, we have the bound

$$\big|\mathbb{E}[Z_{n+1}(x,u)|\mathcal{F}_n]\big| \le e^{-\rho}\max_{x,u}\big|Q_n^\pi(x,u) - Q^\pi(x,u)\big| = e^{-\rho}\|Y_t\|_\infty.$$

We also know (using $(a+b)^2 \le 2a^2+2b^2$) that

$$(Z_{n+1}(x,u))^2 \le 2\Big(G_n - Q^\pi(x,u) + e^{-\rho}Q^\pi(X_{n+},U_{n+})\Big)^2 + 2e^{-\rho}\Big(Q_n^\pi(X_{n+},U_{n+}) - Q^\pi(X_{n+},U_{n+})\Big)^2,$$

and so

$$\mathbb{E}[(Z_{n+1}(x,u))^2|\mathcal{F}_n] \le c(1+\|Y\|_\infty^2)$$

for some $c>0$. Combining with our previous bound, we get the desired growth bound on the variance. As $e^{-\rho}<1$, applying Lemma 2.3.4 we conclude that $\|Y_n\|_\infty \to 0$ a.s., as desired.

3. Consider a modification of the value iteration process, where instead of updating all states simultaneously, we iterate through the states and only update the value associated with each state in turn. Prove that this process gives a convergent approximation of the true value function.

   Proof. Comparing with the proof of convergence given in lectures, we define the operator $\mathcal{T}^i$ which is given by

   $$(\mathcal{T}^i v)(x) = \begin{cases} v(x) & \text{if } i \neq x \\ (\mathcal{T}v)(x) & \text{if } i = x. \end{cases}$$

   Now it is not the case that $\mathcal{T}^i$ is a contraction, but it is true that

   $$\|\mathcal{T}^i v(x) - \mathcal{T}^i v'(x)\| \leq \begin{cases} \|v - v'\|_\infty & \text{if } i \neq x \\ e^{-\rho}\|v - v'\|_\infty & \text{if } i = x. \end{cases}$$

   In particular, we see that

   $$\begin{aligned} \Big|\prod_{i=1}^{|\mathcal{X}|}\mathcal{T}^i v(x) - \prod_{i=1}^{|\mathcal{X}|}\mathcal{T}^i v'(x)\Big| &= \Big|\mathcal{T}^x\prod_{i<x}\mathcal{T}^i v(x) - \mathcal{T}^x\prod_{i<x}\mathcal{T}^i v'(x)\Big| \\ &\leq e^{-\rho}\Big\|\prod_{i<x}\mathcal{T}^i v - \prod_{i<x}\mathcal{T}^i v'\Big\|_\infty \\ &\leq e^{-\rho}\|v - v'\|_\infty \end{aligned}$$

   So the concatenation $\prod_{i=1}^{|\mathcal{X}|}\mathcal{T}^i$ is a contraction, which gives our result. □

4. Consider the Markov Decision problem where an agent wishes to minimize

   $$J(U) = \mathbb{E}^U\Big[\sum_{t=0}^{\tau} g(X_t, U_t)\Big]$$

   where τ is a geometric random variable independent of the control and the state process (and the other terms are as we usually define them). Show that this is equivalent to a discounted control problem.

   Solution: We know that $\mathbb{P}(\tau = T) = (1-p)p^T$, for some $p > 0$. Therefore, as τ is

independent of the control and X, we can write

$$\begin{aligned}
J(U) &= \sum_T \mathbb{E}^U\Big[\sum_{t=0}^{T} g(X_t, U_t)\Big]\mathbb{P}(\tau = T) \\
&= \mathbb{E}^U\Big[\sum_T \sum_{t=0}^{T} g(X_t, U_t)\Big](1-p)p^T \\
&= \mathbb{E}^U\Big[\sum_t \sum_{T}^{\infty} p^T g(X_t, U_t)\Big](1-p) \\
&= \mathbb{E}^U\Big[\sum_t \Big(\sum_{T}^{\infty} p^T(1-p)\Big) g(X_t, U_t)\Big] \\
&= \mathbb{E}^U\Big[\sum_t p^T g(X_t, U_t)\Big].
\end{aligned}$$

Setting $\rho = -\log p$ gives our usual notation.

Sam Cohen: cohens@maths.ox.ac.uk

Section B

5. For the situation in Question 1, implement value iteration, and compare the convergence of this algorithm with policy iteration.

   Solution: See attached file `PS2_value_policy.ipynb`

6. Consider the pollution regulation problem from Sheet 1. Implement a Q-learning algorithm which solves the problem, and compare the results with the explicit solution for appropriate choices of parameter values illustrating two different optimal solutions.

   Solution: See attached file `PS2_QL.ipynb`

7. In this question, we consider at an agent who faces a discrete time MDP over an infinite horizon, with a time-homogenous cost $g(x,u)$ and transition probabilities $p(x';x,u)$. As usual, our actions take values in a compact set $\mathcal{U}$, and g and p are both continuous with respect to u.

   However, this agent wants to minimize the long-run average cost

   $$\bar{J}(X_0,U) := \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}^U\Big[\sum_{t=0}^{T} g(X_t,U_t)\Big].$$

   We consider this through an approximate infinite-horizon discounted problem, where the agent seeks to minimize

   $$J^\rho(X_0,U) = \mathbb{E}^U\Big[\sum_{t=0}^{T} e^{-\rho t} g(X_t,U_t)\Big].$$

   We write v^ρ for the value function for the discounted problem. We aim to show that taking $\rho \to 0$ gives a problem which converges, in an appropriate sense, to the long-run average cost problem.

   We suppose the following *geometric ergodicity* property holds:

   ***Assumption**: There exist constants $R,\gamma > 0$ such that,*

   $$\max_{U\in\mathbb{U}} \max_{x,\tilde{x},x'\in\mathcal{X}} \Big|\mathbb{P}^U[X_t = x'|X_0 = x] - \mathbb{P}^U[X_t = x'|X_0 = \tilde{x}]\Big| \le Re^{-\gamma t} \quad \textit{for all } t.$$

   This assumption guarantees that, for every feedback control $U_t = u(X_t)$, the state X is a Markov chain under $\mathbb{P}^U$ with a unique stationary distribution π^U, and the distribution of X_t converges to this stationary distribution (geometrically quickly, in total variation norm).

   For simplicity, we write $C = \max_{x,u} |g(x,u)|$.

(a) Assuming there is an optimal control which is of feedback form, show that the long-run average cost criterion is the same as minimizing the expected cost

$$\sum_x \pi^U(x) g(x, u(x))$$

where π^U is the stationary distribution when using the feedback control U.

(*Hint: It may help to know that if a_n is a convergent sequence as $n \to \infty$, then the Cesàro sums $n^{-1}\sum_{m\le n} a_m$ are also convergent as $n \to \infty$, to the same limit.*)

(b) By taking the trivial example $g(x, u) \equiv 1$, or otherwise, explain why $v^\rho(x)$ typically diverges as $\rho \to 0$.

(c) Show that $\tilde{v}^\rho(x) := v^\rho(x) - v^\rho(0)$ satisfies

$$\tilde{v}^\rho(x) = \min_u \Big\{ g(x, u) - (1 - e^{-\rho})v^\rho(0) + e^{-\rho} \sum_{x'} p(x'; x, u)\tilde{v}(x') \Big\}$$

and taking the arg min in this equation gives the same (optimal) controls as v^ρ.

(d) Using the geometric ergodicity assumption, show that

$$|\tilde{v}^\rho(x)| = |v^\rho(x) - v^\rho(0)| \le \frac{R|\mathcal{X}|}{1 - e^{-(\gamma+\rho)}} C.$$

Hint: Fix an optimal policy, and write out the value function as an infinite sum.

(e) Show that $(1 - e^{-\rho})|v^\rho(0)| \le C$.

(f) Using the inequalities above, show that we can take a sequence $\rho \to 0$ such that $\tilde{v}^\rho(x)$ converges to some $\bar{v}(x)$ for every x, and $(1-e^{-\rho})v^\rho(0)$ converges to a constant λ, where $\bar{v}$ and λ satisfy the *ergodic Bellman equation*:

$$\bar{v}(x) = \min_u \Big\{ g(x, u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\}.$$

(g) Show that the λ you have just constructed is the optimum value for the long-run average cost criterion, and a time-homogenous feedback control u is optimal if and only if

$$u(x) \in \arg\min_{\mathcal{U}} \Big\{ g(x, u) - \lambda + \sum_{x'} p(x'; x, u)\bar{v}(x') \Big\}.$$

(*You may assume that there exists an optimal control of feedback form.*)

(h) Show that this problem generally admits other (time dependent) optimal controls.

Solution:

(a) If u is a feedback control, we know that $\mathbb{E}^U[g(X_t, u(X_t))] \to \sum_t \pi^U(x)g(x, u(x))$ as $t \to \infty$ (because of the convergence of the law of X_t). The Cesàro sum $T^{-1}\sum_{t=0}^{T} \mathbb{E}^U[g(X_t, u(X_t))]$ has the same limit as $\mathbb{E}^U[g(X_t, u(X_t))]$ (as it is convergent), giving the result.

(b) In this trivial example, we know that $v^\rho(x) = \sum_{t=0}^{\infty} e^{-\rho t} = \frac{1}{1-e^{-\rho}}$. Taking $\rho \to 0$ we see this diverges. This is the typical behaviour, as the total undiscounted cost becomes infinite when we consider it over an infinite horizon, as we will face the same costs infinitely many times.

(c) This is similar to the deterministic shifts we saw in sheet 1. In particular, we know that, with $\mathcal{T}^\rho$ the usual Bellman operator with discount rate ρ, we have

$$\mathcal{T}^\rho(\tilde{v}) = \mathcal{T}^\rho(v^\rho) - e^{-\rho}v^\rho(0) = v^\rho - e^{-\rho}v^\rho(0) = \tilde{v} + (1 - e^{-\rho})v^\rho(0).$$

Rearrangement gives the desired result. That the optimal control remains unchanged is simply because adding a constant doesn't ever change your optimal controls (as we saw on sheet 1).

(d) If u^ρ is an optimal control for the problem with discount rate ρ, we can write our value function as

$$v^\rho(x) = \sum_t \sum_{x'} p_{0,t}(x'; x, u^\rho)e^{-\rho t}g(x', u^\rho(x'))$$

Subtracting, we see that

$$\begin{aligned}
|v^\rho(x) - v^\rho(0)| &= \Big| \sum_t \sum_{x'} \Big(p_{0,t}(x'; x, u^\rho) - p_{0,t}(x'; 0, u^\rho)\Big)e^{-\rho t}g(x', u^\rho(x'))\Big| \\
&\leq \sum_t \sum_{x'} \Big|p_{0,t}(x'; x, u^\rho) - p_{0,t}(x'; 0, u^\rho)\Big|e^{-\rho t}|g(x', u^\rho(x'))| \\
&\leq \sum_t \sum_{x'} Re^{-\gamma t}e^{-\rho t}C \\
&\leq \frac{R|\mathcal{X}|}{1 - e^{-(\gamma+\rho)}}C.
\end{aligned}$$

(e) Using the same expansion as above, we see

$$|v^\rho(x)| \leq \sum_t \sum_{x'} p_{0,t}(x'; x, u^\rho)e^{-\rho t} \max_{x,u} |g(x, u)| = \frac{1}{1 - e^{-\rho}}C.$$

(f) We notice from the above inequalities that $\tilde{v}(x)$ and $(1 - e^{-\rho})v^\rho(0)$ live in compact sets. Therefore, we can take any sequence $\rho \to 0$, and find a subsequence for

which these terms all converge. Taking limits in the Bellman equation, with $\lambda = \lim(1-e^{-\rho})v^\rho(0)$ and $\bar{v}(x) = \lim \tilde{v}^\rho(x)$, we have

$$\bar{v}(x) = \min_u \Big\{ g(x,u) - \lambda + \sum_{x'} p(x';x,u)\bar{v}(x') \Big\}.$$

(g) From the ergodicity assumption, we know that any control u makes X into a Markov chain with stationary distribution π^U, which satisfies $\pi^U(x') = \sum_x p(x';x,u(x))\pi^U(x)$. Therefore, we know

$$\begin{aligned}
\sum_x \pi^U(x)\bar{v}(x) &= \sum_x \pi^U(x) \min_u \Big\{ g(x,u) - \lambda + \sum_{x'} p(x';x,u)\bar{v}(x') \Big\} \\
&\leq \sum_x \Big(\pi^U(x) g(x,u(x)) \Big) - \lambda + \sum_{x,x'} \Big(\pi^U(x) p(x';x,u(x)) \bar{v}(x') \Big) \\
&= \sum_x \Big(\pi^U(x) g(x,u(x)) \Big) - \lambda + \sum_{x'} \Big(\pi^U(x') \bar{v}(x') \Big)
\end{aligned}$$

and hence

$$\lambda \leq \sum_x \Big(\pi^U(x) g(x,u(x)) \Big).$$

From part (a), this shows that λ is less than or equal to the long-run average cost under the arbitrary feedback control u. We also see that this is an equality if and only if u achieves the minimum in the ergodic Bellman equation, that is, u is an optimal policy.

(h) Simply notice that you can change the control at any finite number of times (arbitrarily) without changing the long-run average cost. Therefore, modifying an optimal control in this way will always yield another optimal control.