Final Honour School of Mathematics Part B

C8.7 Optimal Control
CORRECTED VERSION

2025

Do not turn this page until you are told that you may do so

1. In this question, we will consider a deterministic optimal control problem in continuous time. The problem has a one-dimensional state process $X \in \mathcal{X} = \mathbb{R}$, and control space $\mathcal{U} = (0, \infty)$. The state dynamics are given by $\frac{\mathrm{d}}{\mathrm{d}t} X_t = f(t, X_t, U_t)$, and the (undiscounted) cost is given by $\int_0^T g(t, X_t, U_t)\mathrm{d}t + \Phi(X_T)$, for some terminal time $T > 0$.

   (a) [5 marks] Write down the Hamilton–Jacobi equation which should be satisfied by the value function, along with its boundary conditions, defining all terms used.

   (b) [8 marks] Assuming the Hamilton–Jacobi equation has a C^2 solution, and all other required derivatives exist and are continuous, derive Pontryagin's minimum principle, that is, the system of equations:

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}t} X_t^* &= f(t, X_t^*, U_t^*); & X_0^* &= x_0; \\ \frac{\mathrm{d}}{\mathrm{d}t} q_t &= -\partial_x g(t, X_t^*, U_t^*) - \partial_x f(t, X_t^*, U_t^*) q_t; & q_T &= \partial_x \Phi(X_T^*); \\ U_t^* &\in \arg\min_{u \in \mathcal{U}} \left\{ g(t, X_t^*, u) + f(t, X_t^*, u) q_t \right\} \end{aligned}$$

   and explain how this relates to the optimal control.

   (c) [8 marks] Suppose our dynamics and costs are given by

$$\begin{aligned} f(t, x, u) &= 2e^{-x} - \frac{1}{u} \\ g(t, x, u) &= -\frac{x}{u} + \frac{\ln(1+t)}{u} + \frac{1-t}{(1+t)^2} u + \frac{(\alpha + \beta t)x}{1+t} \\ \Phi(x) &= (1+T)e^{-x} - Tx \end{aligned}$$

   for some constants α and β. Suppose also that $x_0 = 0$ and $T \leqslant 1$. Show that the constants α and β can be chosen such that $X_t^* = \ln(1+t)$ is a locally optimal trajectory.

   (d) [4 marks] Show that, when $T > 1$, the candidate control you have constructed is not optimal, and describe the behaviour of the optimally controlled state, given the value of X_1.

2. Consider a one-dimensional stochastic control problem. We have a controlled process X with dynamics

$$\mathrm{d}X_t = f(t, X_t, U_t)\mathrm{d}t + \sigma(t, X_t, U_t)\mathrm{d}W_t; \qquad X_0 = x_0 \in \mathbb{R}$$

for W a Brownian motion, and our aim is to minimize, among controls taking values in $\mathcal{U}$, the undiscounted cost

$$\int_0^T g(t, X_t, U_t)\mathrm{d}t + \Phi(X_T).$$

(a) [5 marks] Write down the Hamilton–Jacobi–Bellman equation, including the domain and appropriate boundary conditions.

(b) [8 marks] Assuming $w : [0, T] \times \mathbb{R} \to \mathbb{R}$ is a ($C^{1,2}$, polynomial growth) solution to the Hamilton–Jacobi–Bellman equation, prove that $w(0, x_0)$ is less than or equal to the minimal cost of our problem.

You may assume that the stochastic integral against a martingale is always a martingale, and may use Itô's lemma: for any $C^{1,2}$ function b, with X satisfying the dynamics above,

$$b(t, X_t) = b(0, 0) + \int_0^t \partial_t b(t, X_t) + f(t, X_t, U_t)\partial_x b(t, X_t)\mathrm{d}t + \int_0^t \sigma(t, X_t, U_t)\partial_x b(t, X_t)\mathrm{d}W_t + \frac{1}{2}\int_0^t \big(\sigma(t, X_t, U_t)\big)^2 \partial^2_{xx} b(t, X_t)\mathrm{d}t.$$

(c) [9 marks] Consider the setting where $\mathcal{U} = (0, 1)$,

$$f(t, x, u) = 2xu, \qquad \sigma(t, x, u) = \sqrt{u},$$
$$g(t, x, u) = \Big(\frac{1}{u} - 2\sqrt{2x^2 + 1}\Big)e^{-(1+x^2)}, \qquad \Phi(x) = -\exp(-(1 + x^2)).$$

By using an ansatz $v(t, x) = -h(t)w(x)$, or otherwise, find the value function and optimal control for this problem.

(d) [3 marks] Suppose a colleague tells you that, because the terminal value is minimized far from the origin, the optimal control should be to stay far from the origin, in particular the control should be small when $x \ll 0$ or $x \gg 0$. Give a possible reply discussing the behaviour of the drift and cost functions.

3. A shop wishes to sell a new product to a small group of customers, and needs to determine an appropriate strategy.

   On each day the shop approaches u customers at random (without replacement) from the pool of N customers, and these customers choose to purchase (or not). The cost of approaching u customers is given by γu, for some $\gamma > 0$. Every day the shop produces 10 copies of the product, which is the maximum number they can sell, and faces a cost β for each unsold item. Items cannot be stored overnight.

   The approaches to customers must be initiated before any purchase decisions can be made (so, for example, the shop might approach 20 customers, in the hope that at least 10 customers will wish to purchase.)

   The shop wishes to minimize their discounted long-run costs $J(u)$, with a daily discount factor $e^{-\rho} \in [0, 1)$, over an infinite horizon.

   (a) [10 marks] Suppose customers' behaviour is as follows: any customer who purchased the product in the previous two days will refuse to purchase, otherwise they will agree to purchase.

   (i) Describe an appropriate state variable for the problem.

   (ii) Give an expression for the distribution of the number of items sold each day, when the shop makes u approaches.
   [*Hint: It may help to recall that the number of successes in n trials, when sampling without replacement from a finite population of size N containing K successes, follows a Hypergeometric(N, K, n) distribution, whose probability mass function is* $\mathbb{P}(X = k) = \binom{K}{k}\binom{N-K}{n-k}/\binom{N}{n}$.]

   (iii) Write down the Bellman equation satisfied by the value function for the shop.

   (b) [8 marks] In the setting of part (a),

   (i) Write down the value iteration algorithm for computing the value function, defining all relevant notation, and prove it converges.

   (ii) Explain how value iteration can be used to compute a sequence of strategies. Do these strategies satisfy $J(u_n) = v_n$, and do they converge to an optimal strategy?

   (c) [7 marks] Suppose that the shop does not know customer's behaviour, but instead only knows that a customer's purchases over the past two days will affect their choice whether to purchase.

   (i) Write down an online Q-learning algorithm which the shop could use to find an optimal strategy. Under what conditions does this converge to the true Q function?

   (ii) Discuss what the advantages and disadvantages of Q-learning are, when compared with value iteration.