UNIVERSITY COLLEGE LONDON

EXAMINATION FOR INTERNAL STUDENTS

MODULE CODE : **COMPGI13**

ASSESSMENT PATTERN : **COMPGI13B**

MODULE NAME : **Advanced Topics in Machine Learning**

DATE : **27-May-14**

TIME : **14:30**

TIME ALLOWED : **2 Hours 30 Minutes**

TURN OVER

Advanced Topics in Machine Learning, GI13, 2013/14

Answer any THREE questions. Each question is worth 20 marks. Use separate answer books for PART A and PART B. **Gatsby PhD students only**: answer *either* TWO questions from PART A and ONE question from PART B; *or* ONE question from PART A and TWO questions from PART B.

Marks for each part of each question are indicated in square brackets

Calculators are NOT permitted

Part A: Kernel Methods

1. Define by $\phi(x)$ the feature map for the RKHS $\mathcal{H}$ with positive definite kernel

$$k(x,x') = \left\langle \phi(x), \phi(x') \right\rangle_{\mathcal{H}} = \left\langle k(x,\cdot), k(x',\cdot) \right\rangle_{\mathcal{H}}.$$

Recall the reproducing property: $\forall f(\cdot) \in \mathcal{H}$,

$$\langle f(\cdot), \phi(x) \rangle_{\mathcal{H}} = \langle f(\cdot), k(x,\cdot) \rangle_{\mathcal{H}} = f(x). \tag{1}$$

(we will equivalently use the shorthand $f \in \mathcal{H}$). Define μ_P to be the mean embedding for distribution P, i.e. the RKHS function for which

$$\langle f, \mu_P \rangle = \mathbb{E}_{X\sim P} f(X) = \mathbb{E}_{X\sim P} \langle f, \phi(X) \rangle, \tag{2}$$

for any $f \in \mathcal{H}$, where $X \sim P$ denotes the random variable X drawn from the probability distribution P.

(a) Show by the reproducing property and the definition in Eq. (2) that

$$\mu_P(x) = \mathbb{E}_{X\sim P} k(X,x).$$

[3 marks]

(b) A linear operator $A : \mathcal{H} \to \mathbb{R}$ is said to be bounded when

$$|Af| \leq \lambda_A \|f\|_{\mathcal{H}}$$

for all $f \in \mathcal{H}$. Show that if the kernel is bounded, $|k(x,x')| < K$ for some strictly positive $K < \infty$ and all x, x', then the operator

$$A_P f := \mathbb{E}_{X \sim P} f(X)$$

is bounded. Give the corresponding λ_{A_P} as a function of the upper bound K. When an operator is bounded, then by the Riesz representer theorem, there exists an element $\mu_A \in \mathcal{H}$ such that

$$Af = \langle \mu_A, f \rangle .$$

Hence, show that there exists some $\mu_P \in \mathcal{H}$ such that Eq. (2) holds. You may need the Cauchy-Schwarz inequality,

$$|\langle a, b \rangle| \leq \|a\| \|b\|, \tag{3}$$

and a result following from Jensen's inequality,

$$|\mathbb{E}_{X \sim P} f(X)| \leq \mathbb{E}_{X \sim P} |f(X)| .$$

[4 marks]

(c) An algorithm known as Kernel Herding implements the iteration: from time T to $T+1$, make the following two updates:

$$x_{T+1} = \operatorname*{argmax}_{x \in \mathcal{X}} \langle w_T, \phi(x) \rangle \tag{4}$$

$$w_{T+1} = w_T + \mu_P - \phi(x_{T+1}) \tag{5}$$

We run herding for T steps, to obtain the points $x_1, \ldots, x_T$. Assume $k(x,x) = K$ constant across all values of x. Initialise the Herding algorithm with $w_0 = \mu_P$. Show that Herding greedily chooses x_{T+1} to minimize the error in constructing the mean embedding,

$$\mathcal{E}_{T+1} := \left\| \mu_P - \frac{1}{T+1} \sum_{t=1}^{T+1} \phi(x_t) \right\|^2 . \tag{6}$$

Hints: first, expand out Eq. (6), and note that the optimization applies only to terms containing x_{T+1}. Then substitute the current expression for w_T into Eq. (4). You will also need the solution to part 1 of the question.

[5 marks]

(d) Define

$$\hat{P}_T := \frac{1}{T}\sum_{t=1}^{T} \delta_{x_t},$$

where δ_{x_t} denotes the Dirac measure at x_t, meaning that the expectation of a function $f(x)$ under $\hat{P}_T$ is

$$\mathbb{E}_{X\sim\hat{P}_T} f(x) = \frac{1}{T}\sum_{t=1}^{T} f(x_t)$$

We begin the Herding process with some $w_0 \in \mathcal{H}$, which need not be μ_P. *Assume* that $\|w_T\|$ is bounded, i.e. $\|w_T\| \leq C$. Given this assumption, show that for all $f \in \mathcal{H}$,

$$\left|\mathbb{E}_{X\sim P} f(X) - \mathbb{E}_{X\sim\hat{P}_T} f(X)\right| = O(T^{-1}).$$

In other words, if the assumption $\|w_T\| \leq C$ holds, then using "pseudo-samples" x_t obtained by Herding results in a fast decrease in error estimates when taking expectations of RKHS functions. Hints: use the expression for w_T. You may need the reproducing property in Eq. (1), the Cauchy-Schwarz inequality in Eq. (3), and the reverse triangle inequality,

$$\|a\| - \|b\| \leq \|a+b\|.$$

[5 marks]

(e) We run the Herding algorithm using the feature map

$$\phi(x) = \begin{bmatrix} x \\ x^2 \end{bmatrix}.$$

Our goal is to find "pseudo-samples" x_t that are representative of the distribution which generated the samples in figure 1. Explain qualitatively what you think will happen - a diagram might be helpful. How could you do better?

[3 marks]

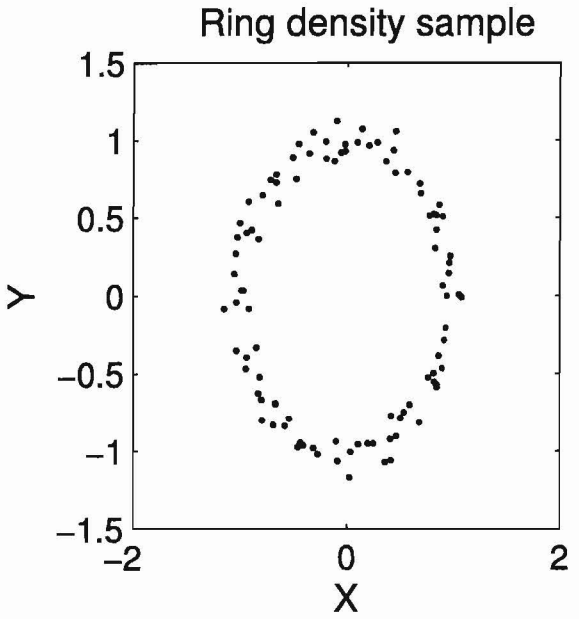

Figure 1: Samples from the "ring density".

2. (a) A symmetric function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is positive definite if $\forall n \geq 1$, $\forall (a_1, \ldots a_n) \in \mathbb{R}^n$, $\forall (x_1, \ldots, x_n) \in \mathcal{X}^n$,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j k(x_i, x_j) \geq 0. \tag{7}$$

Define $\mathcal{F}_1$ to be a reproducing kernel Hilbert space on domain $\mathcal{X}$ with kernel $k_1(x,x')$, and $\mathcal{F}_2$ to be a reproducing kernel Hilbert space on domain $\mathcal{X}$ with kernel $k_2(x,x')$. Show using Eq. (7) that the sum of kernels $k_1(x.x') + k_2(x,x')$ is a kernel. What might be the problem if you took the difference of two kernels - would it still be a kernel? Hint: what happens when you use a single x_1 in Eq. (7)?

[3 marks]

(b) Define $\mathcal{F}$ to be a reproducing kernel Hilbert space on domain $\mathcal{X}$ with feature map $\phi(x)$ and kernel $k(x,x') := \langle \phi(x), \phi(x') \rangle_{\mathcal{F}}$, and $\mathcal{G}$ to be a reproducing kernel Hilbert space on domain $\mathcal{Y}$ with feature map $\psi(y)$ and kernel $l(y,y') := \langle \psi(y), \psi(y') \rangle_{\mathcal{G}}$. Show that the product of the two kernels $k(x,x')$ and $l(y,y')$ is a kernel. Hint: define the tensor product $(a \otimes b)$ between RHKS elements $b \in \mathcal{G}$ and $a \in \mathcal{F}$ such that for all $g \in \mathcal{G}$,

$$(a \otimes b) g \mapsto \langle b, g \rangle_{\mathcal{G}} a.$$

The tensor products between feature maps, $\phi(x) \otimes \psi(y)$, is an element of a Hilbert

space $\mathrm{HS}(\mathcal{G},\mathcal{F})$, with inner product

$$\langle L, M\rangle_{\mathrm{HS}} = \sum_{i\in I} \langle Le_i, Me_i\rangle_{\mathcal{G}}, \tag{8}$$

where $\{e_i\}_{i\in I}$ form an orthonormal basis for $\mathcal{G}$. Given $L = \phi(x)\otimes\psi(y)$ and $M = \phi(x')\otimes\psi(y')$, can you write the above inner product in terms of kernels $k(x,x')$ and $l(y,y')$? Hint: you may use

$$\big\langle \psi(y), \psi(y')\big\rangle_{\mathcal{G}} = \sum_i \langle \psi(y), e_i\rangle_{\mathcal{G}} \big\langle \psi(y'), e_i\big\rangle_{\mathcal{G}}.$$

[3 marks]

(c) Using the previous result, prove that the Gaussian is a kernel,

$$k(x,x') = \exp\left(\frac{-\|x-x'\|^2}{\sigma}\right),$$

using the knowledge that the following function is a kernel,

$$\kappa(x,x') = \exp\left(\frac{\langle x,x'\rangle}{\sigma}\right).$$

[3 marks]

(d) We define a domain $\mathcal{X} := [\ -\pi \quad \pi\ ]$ with periodic boundary conditions. The Fourier series representation of a function $f(x)$ on $\mathcal{X}$ is written $\hat{f}_l$, where

$$f(x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \exp(\imath l x) = \sum_{l=-\infty}^{\infty} \hat{f}_l \left(\cos(lx) + \imath \sin(lx)\right),$$

where $\imath = \sqrt{-1}$. Given a complex number z has the representation $z = A\exp(\imath\theta)$, where θ and A are both real-valued, we define the complex conjugate as $\bar{z} = A\exp(-\imath\theta)$. Assume the kernel takes a single argument, which is the difference in its inputs,

$$k(x,y) = k(x-y),$$

and define the Fourier series representation of k as

$$k(u) = \sum_{l=-\infty}^{\infty} \hat{k}_l \exp\left(\imath l u\right),$$

where we specify $\hat{k}_{-l} = \hat{k}_l$, $\hat{k}_l \geq 0$, and $\overline{\hat{k}_l} = \hat{k}_l$. Define the RKHS inner product as

$$\langle f, g \rangle_{\mathcal{H}} = \sum_{\ell=-\infty}^{\infty} \frac{\hat{f}_\ell \overline{\hat{g}_\ell}}{\hat{k}_\ell}. \tag{9}$$

We represent an RKHS function as

$$f(\cdot) = \left[\begin{array}{ccc} \ldots & \hat{f}_\ell / \sqrt{\hat{k}_\ell} & \ldots \end{array} \right]^\top,$$

and the feature map of a point as

$$k(\cdot, x) = \phi(x) = \left[\begin{array}{ccc} \ldots & \sqrt{\hat{k}_\ell} \exp(-\imath \ell x) & \ldots \end{array} \right]^\top.$$

Show the reproducing property holds,

$$\langle f, k(\cdot, x) \rangle_{\mathcal{H}} = f(x),$$

and that

$$\langle k(\cdot, x), k(\cdot, x') \rangle_{\mathcal{H}} = k(x, x').$$

[4 marks]

(e) Write the Fourier series coefficients of a probability distribution P as $\gamma_{P,\ell}$, and the Fourier coefficients of a distribution Q as $\gamma_{Q,\ell}$ (assume both sets of Fourier coefficients are real-valued). A Fourier series representation of the maximum mean discrepancy is

$$MMD = \left\| \sum_{\ell=-\infty}^{\infty} (\gamma_{P,\ell} - \gamma_{Q,\ell}) \hat{k}_\ell \exp(\imath \ell x) \right\|_{\mathcal{F}}^2.$$

Simplify this expression by using the definition of the RKHS norm arising from Eq. (9).

[3 marks]

(f) Consider the densities

$$p(x) = \frac{1}{2\pi},$$
$$q(x) = \frac{1}{2\pi}(1 + 0.5\cos(rx))$$

where r is integer valued. Assume a Gaussian kernel is used. What will happen to MMD as r increases? For full marks, make reference to the answer in the previous part. Hint: the Fourier series for a Gaussian is itself a Gaussian,

$$k(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-x^2}{2\sigma^2}\right), \qquad \hat{k}_\ell = \frac{1}{2\pi} \exp\left(\frac{-\sigma^2 \ell^2}{2}\right).$$

[4 marks]

3. Define by $\phi(x)$ the feature map for the RKHS $\mathcal{H}$ with positive definite kernel $k(x,x') = \langle \phi(x), \phi(x') \rangle_{\mathcal{H}}$. Recall the reproducing property: $\forall f \in \mathcal{H}$,

$$\langle f, \phi(x) \rangle_{\mathcal{H}} = f(x). \tag{10}$$

(a) Prove that if the the kernel is bounded in absolute value,

$$\left| k(x,x') \right| \leq R \tag{11}$$

then the feature map is bounded, $\|\phi(x)\|^2_{\mathcal{H}} \leq R$.

[2 marks]

(b) We are given a sample

$$S := \{X_1, \ldots, X_m\} \tag{12}$$

drawn i.i.d. from a distribution P. Denote by X a random variable with distribution P. The empirical feature mean as a function of S is

$$\hat{\mu}_P := \frac{1}{m} \sum_{i=1}^{m} \phi(X_i), \tag{13}$$

and by the reproducing property, for all $f \in \mathcal{H}$,

$$\mathbb{E}_{X \sim P} f(X) := \left\langle f, \frac{1}{m} \sum_{i=1}^{m} \phi(X_i) \right\rangle_{\mathcal{H}} = \frac{1}{m} \sum_{i=1}^{m} f(X_i).$$

Denote by μ_P the population feature mean, such that

$$\langle f, \mu_P \rangle = \mathbb{E}_{X \sim P} f(X). \tag{14}$$

You may use without proof the result:

$$\mu_P(x) = \mathbb{E}_{X \sim P} k(X,x) \tag{15}$$

Show via the above two formulae that

$$\langle \mu_P, \mu_P \rangle_{\mathcal{H}} = \mathbb{E}_{X,X'} k(X,X')$$

where X' is a random variable independent of X, but with the same distribution P. Show via the reproducing property and Eq. (14) that

$$\mathbb{E}_{X^m} \langle \mu_P, \hat{\mu}_P \rangle_{\mathcal{H}} = \mathbb{E}_{X,X'} k(X,X').$$

where $\mathbb{E}_{X^m}$ indicates we take the expectation over the m random variables in Eq. (13).

[4 marks]

(c) We define a function of the sample in Eq. (12),

$$g(S) := \left\| \frac{1}{m} \sum_{i=1}^{m} \phi(X_i) - \mu_P \right\|_{\mathcal{H}}$$

which is the expected difference between the sample feature mean and the population feature mean. Show that this expectation satisfies

$$\mathbb{E}_{X^m} g(S) \leq m^{-1/2} \sqrt{\mathbb{E}_X k(X,X) - \mathbb{E}_{X,X'} k(X,X')}$$

Recall Jensen's inequality: for concave functions g,

$$\mathbb{E}_{X \sim P} g(X) \leq g\left(\mathbb{E}_{X \sim P} X\right).$$

[5 marks]

(d) Prove that the previous result can be simplified under the assumption in Eq. (11) to obtain

$$\sqrt{\mathbb{E}_X k(X,X) - \mathbb{E}_{X,X'} k(X,X')} \leq \sqrt{2R}.$$

Hint: you may need the Cauchy-Schwarz inequality,

$$|\langle a, b \rangle| \leq \|a\| \|b\|. \tag{16}$$

[4 marks]

(e) A variance operator $C_{XX} : \mathcal{H} \to \mathcal{H}$ is defined in terms of a probability distribution P_X as

$$\langle f_1, C_{XX} f_2 \rangle_{\mathcal{H}} = \mathbb{E}_X f_1(X) f_2(X) - \mathbb{E}_X f_1(X) \mathbb{E}_X f_2(X).$$

The trace of this operator is written

$$\mathrm{tr}(C_{XX}) := \sum_{i=1}^{\infty} \langle e_i, C_{XX} e_i \rangle_{\mathcal{H}},$$

where $\{e_i\}_{i=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$. Give an expression for the trace in terms of expectations of the kernel. Hint: you will need Parseval's identity,

$$\|h\|_{\mathcal{H}}^2 = \sum_{i=1}^{\infty} \langle h, e_i \rangle_{\mathcal{H}}^2,$$

and a result from earlier in this question (which result to use should be clear from the proof).

Algorithm 1 Greedy Bear

initialize: $Q(a,s) = 0$ for all $s \in \mathcal{S}$, $a \in \mathcal{A}$; initialize s_1; initialize $k = 1$
Choose a_1 greedily from Q
repeat (for each step of episode)
    Observe r_k, s_{k+1}
    Choose a_{k+1} greedily from Q
    $Q(s_k,a_k) \leftarrow Q(s_k,a_k) + \alpha(r_k + \gamma Q(s_{k+1},a_{k+1}) - Q(s_k,a_k))$
    $k \leftarrow k+1$
until episode ends

Part B: Reinforcement Learning

4. A bear can be in one of three places looking for food: the *forest floor*, in the *trees* or by the *river*. From the forest floor he can *climb* into the trees, with no reward, or *forage* for fruit in the bushes for a reward of 1, after which he finds himself on the forest floor with probability $\frac{1}{2}$, or by the river with probability $\frac{1}{2}$. At the river he can *fish* with a reward of 1, after which he remains by the river, or he can *return* to the forest floor with no reward. From the trees he can *descend* back to the forest floor, with no reward, or he can *eat honey* for a reward of 10 but after which he remains in the trees with probability $\frac{1}{2}$ or falls to the forest floor with probability $\frac{1}{2}$.

   (a) Formalise this situation as an MDP with infinite horizon and discount factor $\gamma = 0.5$. Draw a diagram relating states, transitions and rewards for the bear MDP.

   [5 marks]

   (b) (i) Write down the general *Bellman optimality equation* for state-value functions V^*, relating V^* to r, γ, $P^a_{s,s'}$. For the bear MDP use this equation to verify that $V^*(forest) = 8$, $V^*(river) = 4$, $V^*(trees) = 16$, defines the optimal value function.

   [3 marks]

   (ii) Explain how to use the optimal value function to derive an optimal policy π^*. Write down an optimal policy for the bear MDP.

   [3 marks]

The bear does not know the transition dynamics or reward function of its environment and uses the following *Greedy Bear* algorithm to determine its behaviour, with a chosen parameter $\alpha \in (0,1]$. For a given action value function Q we define the greedy policy $\pi'(s) = \text{argmax}_a Q(s,a)$, choosing the maximum uniformly at random in the case of a tie.

(c) The bear observes a trajectory beginning $s_1 = forestfloor, a_1 = forage, r_1 = 1, s_2 = river, \ldots$.

(i) Is the Greedy Bear algorithm model-free or model-based?

[1 marks]

(ii) Perform the first update to the Q function using the Greedy Bear algorithm with $\alpha = 0.5$ to obtain the new Q function.

[1 marks]

(iii) Explain why $Q(forestfloor, forage) > 0$ for the remainder of the algorithm. Will the bear learn an optimal policy using this algorithm? Explain.

[2 marks]

(d) This situation can be improved by using an ε-greedy policy in place of the greedy policy: $\pi(s,a) = \begin{cases} \varepsilon/m + 1 - \varepsilon & \text{if } a = \text{argmax}\, Q(s,a) \\ \varepsilon/m & \text{otherwise} \end{cases}$ where m is the number of actions available in state s. For any ε-greedy policy π, consider the ε-greedy policy π' with respect to Q^π.

(i) Which well-known RL algorithm corresponds to Greedy Bear with the ε-greedy policy?

[1 marks]

(ii) Prove that $Q^\pi(s, \pi'(s)) \geq V^\pi(s)$ for all s.

[4 marks]

[Total 20 marks]

5. In this question we consider multi-armed bandits with a set $\mathcal{A}$ of m arms. When arm a is pulled a reward r is received with probability $R^a(r)$. We aim to maximise cumulative reward. We denote the action-value $Q(a) = \mathbb{E}[r|a]$, and $V^* = \max_{a\in\mathcal{A}} Q(a)$ is the optimal value. We denote the history of experience $h_t = a_1, r_1, a_2, r_2, ..., a_t, r_t$, and let $N_t(a)$ count the number of selections of action a in h_t.

(a) Consider an estimate $\hat{Q}$ for Q.

(i) Define the greedy method with respect to $\hat{Q}$ for selecting arm a and discuss one key problem with the method.

[2 marks]

(ii) Define the ε-greedy method to select the next arm a. Derive a lower bound for the regret $\ell_t = \mathbb{E}[V^* - Q(a_t)]$ at time step t when using the ε-greedy method. How does total regret $\sum_{t=1}^{T} \ell_t$ grow with respect to T?

[4 marks]

(iii) Discuss one possible modification to the ε-greedy method which might reduce total asymptotic regret.

[1 marks]

(b) Let $X_1, ..., X_t$ be iid random variables in $[0,1]$ and let $\bar{X}_t = \frac{1}{t}\sum_{\tau=1}^{t} X_\tau$ denote the sample mean. Recall Hoeffding's inequality,

$$\mathbb{P}(\mathbb{E}[X] > \bar{X}_t + u) < e^{-2u^2 t}.$$

Consider estimating $Q(a)$ from h_t using the estimate

$$\hat{Q}_t(a) = \frac{1}{N_t(a)} \sum_{\tau=1}^{t} 1_{\{a_\tau = a\}} r_\tau.$$

(i) Specialize Hoeffding's bound to the bandit setting to prove an upper confidence bound on the action-value $Q(a)$. For an arbitrary probability p, you should write your bound in the form $\mathbb{P}(Q(a) > A) < p$, where A is defined in terms of the estimate $\hat{Q}_t(a)$, p, and $N_t(a)$.

[3 marks]

(ii) By selecting $p = t^{-4}$, suggest a UCB algorithm based on your upper confidence bound.

[2 marks]

(c) (i) Consider now a 2-armed bandit, $\mathcal{A} = \{A, B\}$. Suppose the history $a_1 = A, r_1 = 3, a_2 = B, r_2 = 1, a_3 = A, r_3 = -1, a_4 = B, r_4 = 1, a_5 = A, r_5 = 1$ has been observed. Compute the estimates $\hat{Q}_5(A)$ and $\hat{Q}_5(B)$. Explain which action your UCB algorithm would select at round 6.

[2 marks]

(ii) Considering multi-armed bandits with a finite number of arms and with bounded rewards $r \in [0, 1]$, is there an upper bound on the number of times your UCB algorithm will pull any arm? Explain.

[3 marks]

(iii) Describe a forward search algorithm for Markov Decision Processes that makes use of the UCB algorithm at every step (do not use random policies at any point). Give pseudocode, or an informal description of your algorithm outlining how this algorithm would find the optimal action from a root state s_1.

[3 marks]

[Total 20 marks]

6. In this question the principal of dynamic programming will be applied to the problem of linear-quadratic control. Consider the case where the planning horizon is finite, $T \in \mathbb{N}$, and both the transition dynamics and reward function are deterministic.

In this problem the state and action are one-dimensional continuous variables, with $s \in \mathbb{R}$ and $a \in \mathbb{R}$. The transition dynamics are linear,

$$s_{t+1} = As_t + Ba_t,$$

the reward function is quadratic,

$$R(s_t, a_t) = Us_t^2 + Va_t^2,$$

and we have $A, B \in \mathbb{R}$ and $U, V \in \mathbb{R}^-$.

For each $t \in \mathbb{N}_T$ define the action-value function, $Q_t(s_t, a_{t:T})$, as the total reward from state s_t onwards, when taking the sequence of actions $a_{t:T} = (a_t, a_{t+1}, ..., a_T)$,

$$Q_t(s_t, a_{t:T}) = \sum_{\tau=t}^{T} R(s_\tau, a_\tau).$$

Note that, as the system is completely deterministic, the states $s_{t+1:T} = (s_{t+1}, ..., s_T)$ are completely determined by s_t and $a_{t:T}$. Given an initial state, s_1, the objective is to find a sequence of actions, $a_{1:T}$, that maximises the total reward, $Q_1(s_1, a_{1:T})$.

(a) For $t \in \mathbb{N}_T$, write the action-value function $Q_t(s_t, a_{t:T})$ directly in terms of U, V, $s_{t:T}$ and $a_{t:T}$.

[4 marks]

(b) Write down the *Bellman equation* for this system, i.e. write down the recursive relationship between $Q_t(s_t, a_{t:T})$ and $Q_{t+1}(s_{t+1}, a_{t+1:T})$, for $t \in \mathbb{N}_{T-1}$.

[3 marks]

(c) Denote the optimal action-value function at time, $t \in \mathbb{N}_T$, by $Q_t^*(s_t)$, i.e.

$$Q_t^*(s_t) = \max_{a_{t:T}} Q_t(s_t, a_{t:T}).$$

Write down the *Bellman optimality equation* for this system, i.e. write down the recursive relationship between $Q_t^*(s_t)$ and $Q_{t+1}^*(s_{t+1})$, for $t \in \mathbb{N}_{T-1}$.

[3 marks]

(d) For $t \in \mathbb{N}_{T-1}$, suppose that $Q^*_{t+1}(s_{t+1})$ can be written in the form

$$Q^*_{t+1}(s_{t+1}) = P_{t+1}s^2_{t+1},$$

for some $P_{t+1} \in \mathbb{R}^-$. Denoting the optimal actions from time $t+1$ to T by $a^*_{t+1:T}$, use this assumption to show that $Q_t(s_t, a_t, a^*_{t+1:T})$ can be written in the form

$$Q_t(s_t, a_t, a^*_{t+1:T}) = \left(U + P_{t+1}A^2\right)s_t^2 + \left(V + P_{t+1}B^2\right)a_t^2 + 2s_tAP_{t+1}Ba_t.$$

[3 marks]

(e) Use this relation to show that the optimal action at time t takes the form

$$a^*_t = -Ks_t,$$

where, $K \in \mathbb{R}$, takes the form

$$K = \left(V + P_{t+1}B^2\right)^{-1}BP_{t+1}A$$

[3 marks]

(f) Use parts (e) and (f) to show that $Q^*_t(s_t)$ takes the form

$$Q^*_t(s_t) = P_ts^2_t,$$

and give an explicit form for P_t in terms of P_{t+1}, A, B, U and V.

[3 marks]

(g) Using your relation for P_t, show that $P_t \leq 0$.

[1 marks]

[Total 20 marks]

END OF PAPER