ECE 555 Second Mid-term Exam Spring, 2011

Good luck!

Full Name: .....

Problem 1:	.....	/ 15 points
Problem 2:	.....	/ 15 points
Problem 3:	.....	/ 20 points
Problem 4:	.....	/ 10 points
Problem 5:	.....	/ 40 points
TOTAL:	.....	/ 100 points

Notes: This is a take home exam.
It is due *5pm on Thursday, May 12, in 162 CSL*

I will make the following times available for discussion:

- Tuesday, May 3rd after lecture
- Monday, May 9th, 3:30-5:00pm
- Wednesday, May 11th, 4:00-5:30pm

Of course, Dayu, Wei, and I will also await your questions via email.

Problem 1 Recall from the first midterm: If f is a real-valued function on $\mathsf{X} = \{0, 1, \dots, 100\}$, then the minimization of f can be expressed as a linear program. Namely, 6+6+3 pts.

$$\min\{\pi(f) : \pi(x) \geq 0,\ x \in \mathsf{X},\ \text{and}\ \sum \pi(x) = 1\}$$

(a) Compute the dual functional,

$$\varphi(z) = \min\{\pi(f) + z\big(1 - \sum \pi(x)\big) : \pi(x) \geq 0,\ x \in \mathsf{X}\}$$

(b) Solve the dual linear program, $\max_{z \in \mathbb{R}} \varphi(z)$

(c) *Compare your conclusions to the duality theory for MDPs derived in class*

Problem 2 Recall the average-cost optimality equation (ACOE): 15 pts.

$$\min_u \big\{ c(x,u) + P_u h^*(x) \big\} = h^*(x) + \eta^* \tag{1}$$

We view this as a fixed point equation in (h^*, η^*), and can apply standard methods for solving such equations. However, recall that h^* is not unique - we can always add a constant to obtain a new solution to (1). To normalize h^* we take $h^*(x^\circ) = \eta^*$, where x° is some distinguished state. On denoting by $T(h)$ the functional defined by $T(h)(x) = \min_u \big\{ c(x,u) + P_u h(x) \big\} - h(x^\circ)$, the ACOE becomes the fixed point equation

$$h^* = T(h^*) \tag{2}$$

Derive the Newton-Raphson algorithm to solve this fixed point equation, and compare your algorithm to PIA.

Problem 3 One student this term voiced alarm: *Average cost optimal control does not take into account the variance of the cost.* A convenient approach to address the impact of risk is the risk-sensitive control problem — A version is illustrated in this problem. 5+5+10 pts.

Let c denote a cost function, and denote the partial sums by $S_0 := 0$, and

$$S_n = \sum_{t=0}^{n-1} c(X(t), U(t)), \qquad n \geq 1$$

For $\theta > 0$ we have the Taylor-series approximation,

$$\exp(\theta S_n) \approx 1 + \theta S_n + \tfrac{1}{2}\theta^2 (S_n)^2$$

This motivates the risk-sensitive value function,

$$h^*(x) = \min \sum_{n=0}^{\infty} \beta^n \mathsf{E}\big[\exp\big(\theta S_n\big) \mid X(0) = x\big]$$

where $\beta \in (0, 1)$, $\theta > 0$ (and typically small). The minimum is over all policies that are adapted to $\boldsymbol{X}$.

(a) Obtain a dynamic programming equation. That is, show that h^* solves the fixed point equation of the form,

$$h^*(x) = \min_u \{\exp(c(x, u)) + \textbf{?}\}$$

where the question-mark depends on x, u, and h^*.

For the remaining questions you may assume that the state space and action space are finite.

(b) Provided h^* is finite valued, the optimal policy is expressed as state feedback, $U(t) = \phi^*(X(t))$. Include a characterization of ϕ^*.

(c) Formulate a VIA algorithm, and derive conditions for convergence.

Problem 4 In this exercise you will see why the original SARSA paper was never published, even though it is a very useful idea. 3+4+3 pts.

Suppose that we have an MDP model (you can assume finite state and action space), and the input $\boldsymbol{U}$ is defined by a randomized, stationary policy ϕ. That is, we have for any x, u, t,

$$\mathsf{P}\{U(t) = u \mid X_0^t,\, X(t) = x\} = \phi_u(x)$$

A cost function c is given, and discount factor $\beta \in (0, 1)$. The value function h for this policy solves the fixed point equation,

$$h(x) = c_\phi(x) + \beta P_\phi h\,(x) \qquad (*)$$

As seen in class on April 28th, on writing $Q(x, u) = c(x, u) + \beta P_u h\,(x)$, and $Q_\phi(x) = \mathsf{E}[Q(X, U) \mid X = x]$, this "Q-function" satisfies a similar equation,

$$Q(x, u) = c(x, u) + \beta P_u Q_\phi\,(x) \qquad (**)$$

In the following steps you will transform (**) into (*).

(a) Explain why $\Phi(t) = (X(t), U(t))$ is a Markov chain (under the assumptions of this exercise).

(b) Obtain an expression for the conditional expectation $\mathsf{E}[Q(\Phi(t + 1)) \mid \Phi(t) = \xi]$ for $\xi = (x, u)$ satisfying $\phi_u(x) > 0$.

(c) Based on your solution to (b), explain why (**) is simply (*) for the chain Φ.

Problem 5 In this exercise you will apply TD-learning in an attempt to solve a particular consensus problem – for background, see the examples in: "Q-learning and Pontryagin's minimum principle". *Proc. of the IEEE Conf. on Dec. and Control*, 2009. 50 pts.

Model: There are $N > 1$ "agents" represented by a scalar state process $\boldsymbol{X}^i$ and scalar input process $\boldsymbol{U}^i$. Each agent evolves according to a linear state space model,

$$X^i(t+1) = \alpha^i X^i(t) + (1-\alpha^i)U^i(t) + \sigma^i W^i(t+1), \quad t \geq 0.$$

where $\boldsymbol{W}^i$ is an i.i.d. sequence with zero mean and unit variance. The processes $\{\boldsymbol{W}^i\}$ are mutually independent. Each agent has access only to its own state and action process $(\boldsymbol{X}^i, \boldsymbol{U}^i)$, and the average $\bar{X}(t) = N^{-1}\sum X^i(t)$. Based on this information, each agent wishes to optimize the average cost,

$$\eta^i = \limsup_{n\to\infty} \frac{1}{n}\sum_{t=1}^{n} c(X^i(t), \bar{X}(t), U(t))$$

where $c(x, \bar{x}, u) = (x-\bar{x})^2 + u^2$.

This is a decentralized optimal control problem that falls outside of the scope of this course. Solutions have been obtained only in the mean field limit: See the lecture by Peter Caines on April 26th, or the bibliography in the aforementioned paper.

In the mean-field limit, we view $\bar{X}$ as a constant. In this case, optimization of η^i by the ith agent amounts to scalar state feedback in the form $U^i(t) = -K_i^*(X^i(t) - \bar{X})$. The relative value function is of the form $h^*(x) = m^*(x-\bar{X})^2$, with $m^* > 0$.

(a) Formulate a two dimensional basis for TD-learning, based on the intuition that $\bar{X}(t)$ is "almost" static, and "almost" independent of any individual action.

(b) How might you construct a Q-learning algorithm with basis?
What would a basis look like, given your answer to (a)?

(c) Describe a TD-learning algorithm to obtain a feedback policy of the form $U^i(t) = \phi_i(X^i(t), \bar{X}(t))$. You will use policy improvement, and you must introduce randomization. That is, if in the kth step of the algorithm you have a value function approximation $h^{k,i}$ for the ith agent, then the policy improvement step would give,

$$\phi_{k,i}(x, \bar{x}) = \arg\min_u \{c(x, \bar{x}, u) + P_u(x, \bar{x})\}$$

where P_u is a *model* for the ith agent, that pretends $(x, \bar{x})$ is the Markov state for a fully-observed MDP.

Note: If you like, you can instead formulate a Q-learning algorithm, or substitute TD with SARSA.

(d) Try out the algorithm with $N = 10$. At the start of your experiment you should choose the $\{\alpha^i, \sigma^i\}$ at random, uniformly on $[.1, .9] \times [0, 2]$, but then keep them fixed thereafter, even in multiple runs.

If you don't like this problem, propose your own "Problem 5" that will exercise the same concepts. **PTO...**

Discussion on Problem 5: This is designed to make you think about all of the concepts you have learned in the course. You won't use all of them, but try to bring in as many ideas as possible.

The first tasks are analytical/philosophical, and the last task is numerical, but I hope you bring in your own ideas. Please explore! Please provide discussion regarding your conclusions!! In particular,

(i) Give me data: Plots, interpretations, histograms, anything you feel will illustrate your findings.

(ii) Does your final policy $\phi_i(x, \bar{x})$ resemble what would be predicted from the infinite-population model, with $\bar{X}$ fixed? Does the value of (α^i, σ^i) influence this comparison?

(iii) Formulate your own questions. Examples: Is exploration needed? Do you find sensitivity of the final outcome to the initial policy?