UC Berkeley
Department of Electrical Engineering and Computer Sciences

EECS 126: Probability and Random Processes

Problem Set 14 (Optional)
Fall 2020

1. Maximum Entropy Value Iteration

In a standard planning problem, we seek to minimize our expected cost $\mathbb{E}_\pi[\sum_{t=1}^{H} c(s_t, \pi(s_t))]$ by optimizing a policy $\pi(s)$ that outputs an optimal action from the state s (possibly also in terms of t). The approach we discussed in lecture was dynamic programming, or Bellman equations; this is also known as value iteration.

Now, consider a *stochastic* policy, where the policy outputs a probability distribution over actions, instead of a deterministic function. One desirable property of making policies stochastic is that they are more robust to errors: if we pre-plan to optimize against c under a certain dynamics function $s_{t+1} = f(s_t, \pi(s_t))$, we may *overfit* to our supposed dynamics, performing poorly in the true dynamics. Furthermore, they allow for better exploration: if we are stuck in a maze and have planned against imperfect dynamics, wandering around randomly may help us find the goal. Stochastic policies are common in modern AI, many of which seek to optimize a similar objective which we explore here.

We can quantify how much we would like our policies to be stochastic vs deterministic by adding an entropy bonus modulated by β, instead optimizing the following expected cost:

$$\mathbb{E}_{u_t \sim \pi(s_t)}[\sum_{t=1}^{H} c(s_t, u_t) - \beta H(\pi(s_t))] \tag{1}$$

where $u_t \sim \pi(s_t)$ and H denotes the entropy of the probability distribution $\pi(s_t)$. Intuitively, for low β we recover a deterministic policy that optimizes against the dynamics completely, whereas for high β we get a uniform policy.

Derive the new value iteration update equations for this modified cost function.

2. Infinite-Horizon Discounted Cost MDP

Consider a relay placement problem, where a deployment agent starts walking from state 0 on a line. He stops at regular intervals (consider the step length to be $\delta > 0$), and decides whether to place a relay there or not. At each step, he measures the power required to maintain a reasonable quality link. The goal of the agent is to minimize a linear combination of power cost and relay cost. Assume that the process restarts every time the agent deploys a relay. Also assume the length of the line is geometric with parameter θ. Formulate the problem as an infinite horizon MDP, with state space (r, γ) where r and γ are the respective distance and power from the previously placed relay. Using the Bellman equation, find the optimal policy structure. Mention a way to compute the optimal policy.

3. Cheapest Fare using HMM

Companies A and B run identical buses in Berkeley, where company A has higher fares. The number of people in buses run by company A and B is a Poisson random variable with rate 10 and 20, respectively. You are counting the number of people on the buses at a

bus-stop where only one bus comes each hour; let X_k be the bus company and N_k be the number of people in the k-th hour, respectively. A Markov chain with transition probabilities $P(X_{k+1} = B|X_k = A) = 0.7$ and $P(X_{k+1} = A|X_k = B) = 0.8$ determines the company of buses arriving.

(a) Let the initial state be $X_1 = A$. What is $P(X_2 = A|N_2 = 13)$?

(b) Assuming that the initial state was $X_1 = A$, say you observed the number of people $N_1 = 7$, $N_2 = 21$ and $N_3 = 9$ in the first three hours. What is your MLSE estimate for the sequence of first three buses?

(c) You board the bus in the fourth hour. Assuming that the most likely sequence is true, what is the probability that you board the bus with a cheaper fare?

4. Higher-Order Markov Chains

Let k be a fixed positive integer. A stochastic process $(X_n)_{n\in\mathbb{N}}$ taking values in a discrete state space $\mathcal{X}$ is called a **kth order (time homogeneous) Markov chain** if for all $n \in \mathbb{N}$ and all feasible sequences $x_0, x_1, \ldots, x_{n+k} \in \mathcal{X}$,

$$\begin{aligned}&\mathbb{P}(X_{n+k} = x_{n+k} \mid X_0 = x_0, X_1 = x_1, \ldots, X_{n+k-1} = x_{n+k-1})\\ &\quad = \mathbb{P}(X_{n+k} = x_{n+k} \mid X_n = x_n, \ldots, X_{n+k-1} = x_{n+k-1})\\ &\quad = P_k(x_{n+k} \mid x_n, \ldots, x_{n+k-1}).\end{aligned}$$

In other words, the transition to the next state depends only on the previous k states. For example, if X_n represents the position of a particle moving with constant velocity at time n, then the system is a second-order Markov chain because the previous two position measurements are needed to infer the particle's velocity.

Show that we can "embed" $(X_n)_{n\in\mathbb{N}}$ into a *first-order* Markov chain $(Z_n)_{n\in\mathbb{N}}$ with an augmented state space, in the sense that X_n can be recovered from Z_n. This allows us to apply algorithms such as the Viterbi algorithm to systems with higher orders of dependence.

5. *Bonus*: EM for a Simple HMM

Consider an HMM $(X_i)_{i\in\mathbb{N}}$ on the state space $\{0, 1\}$, where

$$P(0, 1) = P(1, 0) = \theta \in (0, 1)$$

is an unknown parameter. The hidden state is observed through a BSC with known error probability $\epsilon \in (0, 1)$; let the observations be denoted $(Y_i)_{i\in\mathbb{N}}$. For a fixed positive integer n, suppose that we observe $Y_0, Y_1, \ldots, Y_n$. The initial hidden state is equally likely to be 0 or 1.

(a) What is the MLE for θ given $X_1, \ldots, X_n$? (Use the notation

$$T = \sum_{i=1}^{n} \mathbf{1}\{X_i \neq X_{i-1}\}$$

for the number of times that the hidden state switches between 0 and 1.)

(b) We will now derive an EM algorithm to estimate θ given $Y_1, \ldots, Y_n$. Initialize a guess $\hat{\theta}^{(0)}$. For $t = 0, 1, 2, \ldots$:

- **E step**: Compute $\bar{X}^{(t)} := \mathbb{E}_{\hat{\theta}^{(t)}}[n^{-1}T \mid Y_0, Y_1, \ldots, Y_n]$.
- **M step**: In this case, the next parameter estimate is $\hat{\theta}^{(t+1)} := \bar{X}^{(t)}$.

Explicitly write out what the E step is.