Homework 1

MS&E 338-Winter 2017
due: 01/25/2017

In this assignment we will examine the performance of Thompson sampling and UCB on a linear bandit problem. As your solution, prepare a short report including the figures asked for in the problem along with their descriptions and comparisons. You can return your solution to the class folder located on the second floor of Packard building.

Problem 1. Consider a linear bandit problem where there exist 100 possible actions[1] $\{z_i\}_{i=1}^{100}$, where $z_i \in \mathbb{R}^2$, for $i = 1, 2, \cdots, 100$. Upon playing action z_i, a random reward is generated according to

$$y = \theta^\intercal z_i + w,$$

where $\theta \in \mathbb{R}^2$ is the unknown parameter of the problem and $w \sim N(0, 1)$ is the observation noise which is iid across different time steps and independent of θ. Assume that the prior distribution over θ is a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^2$ and covariance matrix I (the 2×2 identity matrix).

For the following parts, you need to average your results over ten thousands random realizations of the problem. At each realization, you can generate the actions $z_1, z_2, \cdots, z_{100}$ by sampling from a multivariate normal distribution with mean $\mathbf{0} \in \mathbb{R}^2$ and covariance matrix I. We are going to consider the problem on a horizon consisting of $T = 100$ time steps.

(a) Implement Thompson sampling algorithm for the linear bandit problem described above and plot its expected per-period regret.

(b) A version of UCB for the above problem is as follows.

1. At time step t:
   - let $\hat{\theta}_t = \mu_t$ and $\Psi_t = \Sigma_t^{-1}$, where μ_t and Σ_t are the mean and the covariance matrix of the posterior distribution

[1] Note that the action space in a linear bandit problem is usually a convex set with infinitely many actions. For simplicity, we consider a finite action set in this assignment.

- let $\Theta_t = \left\{ \phi : \|\phi - \hat{\theta}_t\|_{\Psi_t} \leq \beta\sqrt{2\log t} \right\}$ be the confidence set, where $\|x\|_\Psi = \sqrt{x^\intercal \Psi x}$
- play the action $i_t = \underset{1 \leq i \leq 100}{\operatorname{argmax}} \; \max_{\phi \in \Theta_t} \phi^\intercal z_i$

2. Update the posterior, increment t and go to step 1

Implement the above version of UCB for various values of β and compare its expected per-period regret with that of Thompson sampling.

(c) Now, consider an agent who is agnostic to the true model of the environment and assumes that the mean rewards of the arms are independent. In this case, the agent would have a misspecified prior distribution of $N(0, 1)$ on the mean reward of each action. Implement UCB (for various values of β) and Thompson sampling algorithms under this misspecified model and compare their performance to the thoughtful versions. Note that in this case, although the agent assumes independent arms, the rewards are still generated according to the linear model described above.