Question 3 Part d

Tuesday, June 4, 2019 12:13 PM

Papers:

1. M. Abeille and A. Lazaric. Thompson Sampling for Linear-Quadratic Control Problems. In AISTATS, 2017.
2. Yasin Abbasi-Yadkori and Csaba Szepesvári. Bayesian optimal control of smoothly parameterized systems. In Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2015

Overview

One exploration technique taken in this paper by Abeille and Lazaric is to use Thompson sampling for LQR learning problems. The problem is divided into a set of episodes. In each episode, the design matrix (matrix of controls and states up to that point) and the regularized least-squares (RLS) estimate of the parameters (θ) are computed. The new RLS estimate $\widetilde{\theta}_t$ is not based on a Bayesian structure or a Gaussian prior assumption but instead, this method samples a perturbed RLS-estimate, while also checking that the new estimate is admissible.

while $\widetilde{\theta}_t \notin \mathcal{S}$ **do**
  Sample $\eta_t \sim \mathcal{D}^{\mathrm{TS}}$
  Compute $\widetilde{\theta}_t = \widehat{\theta}_t + \beta_t(\delta')V_t^{-1/2}\eta_t$
end while

Every coordinate of the matrix η_t is a random sample drawn i.i.d. from $\mathcal{N}(0,1)$.
The overarching idea of this Thompson sampling process is to sample the parameters $\widetilde{\theta}_t$ in each episode such that the action chosen $u_t = K\left(\widetilde{\theta}_t\right)x_t$ will maximize the expected reward given the sampled parameters, the action, and the current history of states and controls.

Pros

The main benefit of Thompson sampling is that computing the Bayesian optimal policy is computationally expensive. This algorithm is a lazy posterior sampling method that maintains a distribution over the unknown parameter and changes the policy only when the variance of the distribution is reduced sufficiently. This allows the algorithm to trade off performance for computational efficiency while still getting near Bayesian optimal results.

Cons

One of the major difficulties with Thompson sampling that is mentioned by the authors is that Thompson sampling has to solve a trade-off between frequently updating the policy to guarantee enough optimistic samples and minimizing the number of policy switches to limit the regret incurred at each change. Solving this tradeoff leads to a complexity bound of $O\left(T^{2/3}\right)$ as compared to Optimism in the Face of Uncertainty approaches that have a smaller complexity bound of $O\left(T^{1/2}\right)$.