B9120-001 Dynamic Programming Fall 2023

Prof. Daniel Russo

Homework Assignment 9: Due Friday December 8

Read Bertsekas Vol II, Section 2.4. and/or the course notes to refresh your understanding of policy iteration.

This problem explores a precise connection between policy iteration and the conditional gradient algorithm (a.k.a Frank Wolfe) applied to the policy gradient objective.

MDP setup: Let $\mathcal{X} = \{1, \cdots, n\}$ and $U = \{u \in \mathbb{R}^k : \mathbf{1}^\top u = 1\, ,\, u \succeq 0\}$ be the set of probability distributions over k base actions. Due to the linearity of expectations, expected costs and transition probabilities are linear in the stochastic action vector, with

$$g(x,u) = \sum_{i=1}^{k} g(x, e_i) u_i \qquad p(x'|x,u) = \sum_{i=1}^{k} p(x'|x, e_i) u_i$$

where e_i is the i-th standard basis vector. The set of stochastic stationary policies $\Pi = \{\pi \in \mathbb{R}^{n \times k} : \pi_x \in U \ \forall x\}$ is the set of matrices whose rows are probability distributions. Define

$$\ell(\pi) = w^\top J_\pi = \sum_{x=1}^{n} w(x) J_\pi(x) \qquad \pi \in \Pi$$

for given state-relevance weights w where $w(x) > 0$ and $\sum_{x=1}^{n} w(x) = 1$.

Next, define the advantage function

$$A_\pi(x,u) = \left(g(x,u) + \sum_{x' \in \mathcal{X}} p(x'|x,u) J_\pi(x') \right) - J_\pi(x),$$

which is the difference in long-term cost between a) applying u in state x and following π thereafter and b) applying π throughout.

Policy iteration: In this notation, policy iteration produces a sequence of iterates $\{\pi_k\}_{k \in \mathbb{N}}$ where

$$\pi_{k+1}(x) \in \arg\min_{u \in U} A_{\pi_k}(x,u) \ \forall x \in \mathcal{X}.$$

Conditional gradient algorithm: Consider the conditional gradient (CG) algorithm applied to $\min_{\pi \in \Pi} \ell(\pi)$. Beginning with some initial iterate π_0, CG produces a sequence of iterates $\{\pi_k\}_{k \in \mathbb{N}}$ where

$$\pi_{k+1} = (1 - \gamma_k)\pi_k + \gamma_k y_k \tag{1}$$

$$y_k \in \arg\min_{\pi \in \Pi} \langle \nabla \ell(\pi_k)\, ,\, \pi - \pi_k \rangle \tag{2}$$

where $\gamma_k \in (0,1)$. We approximate ℓ by linearizion around π_k, minimize that approximation globally (here solving an LP), and then take a small step in that direction (reflecting that the linearization is not globally accurate).

In class, we showed following first order Taylor expansion of the policy gradient objective:

$$\ell(\pi^+) = \ell(\pi) + \sum_{x=1}^{n} d_\pi(x) \underbrace{\left(g(x, \pi_x^+) + \sum_{x' \in \mathcal{X}} p(x'|x, \pi^+(x)) J_\pi(x') - J_\pi(x) \right)}_{=(T_{\pi^+} J_\pi - J_\pi)(x)} + O(\|\pi^+ - \pi\|^2),$$

where $d_\pi(x) = \mathbb{E}[\sum_{k=0}^{\tau-1} 1(x_k = x) \mid x_0 \sim w]$ is the occupancy measure under x. In different notation, one could rewrite this as

$$\ell(\pi^+) = \ell(\pi) + \sum_{x=1}^{n} d_\pi(x) A_\pi(x, \pi_x^+) + O(\|\pi^+ - \pi\|^2). \tag{3}$$

You may assume (3) holds and use it in the subsequent problems.

If you can stuck on a subproblem, you may solve the remaining subproblems assuming its claim.

Part (a) Recognize that $A_\pi(x,u)$ is linear (technically, "affine") in u.

Part (b) Calculate $\frac{\partial}{\partial \pi_{x,u}} \ell(\pi)$.

Part (c) Show that y_k in (2) is a policy iteration update to π_k.

Part (d) Assume the Bellman operator T is a contraction in the supremum norm $\|\cdot\|_\infty$ with modulus γ (as in the discounted case). Consider a fixed stepsize $\gamma_k = \gamma \in (0,1)$. Show that $\|J_{\pi_k} - J^*\|_\infty \leq (1 - \gamma(1-\alpha))^k \|J_{\pi_0} - J^*\|_\infty$. Provide the same convergence result for $\ell(\cdot)$. *Hint: what do you know about policy iteration's convergence rate and the proof of that?*

Part (e) What stepsize choice does your analysis suggest?

You don't need to write anything, but think about how to reconcile (e) with the usual stepsizes in smooth optimization.