CMPT 419/983: Theoretical Foundations of Reinforcement Learning

Lecture 9

Sharan Vaswani

November 3, 2023

Recap

Politex

- **Policy Evaluation**: Compute the estimate $\hat{q}_k := \hat{q}^{\pi_k}$ and define $\bar{q}_k := \sum_{i=0}^{k} \hat{q}_i$.
- **Policy Update**: $\forall (s,a)$, $\pi_{k+1}(a|s) = \frac{\exp(\eta\, \bar{q}_k(s,a))}{\sum_{a'} \exp(\eta\, \bar{q}_k(s,a'))}$.
- If $\hat{q}^k = q^{\pi_k} + \epsilon_k$, $\|v^{\bar{\pi}_K} - v^*\|_\infty \leq \frac{\|\text{Regret}(K)\|_\infty}{(1-\gamma)\, K} + \frac{2 \max_{k \in \{0,\ldots,K-1\}} \|\epsilon_k\|_\infty}{(1-\gamma)}$, where $\text{Regret}(K) = \sum_{k=0}^{K-1} [\mathcal{M}_{\pi^*} \hat{q}_k - \mathcal{M}_{\pi_k} \hat{q}_k] \in \mathbb{R}^S$. $\|\text{Regret}(K))\|_\infty = \max_s |R_K(\pi^*, s)|$, where $R_K(\pi^*, s) := \sum_{k=0}^{K-1} \langle \pi^*(\cdot|s), \hat{q}_k(s,\cdot)\rangle - \langle \pi_k(\cdot|s), \hat{q}_k(s,\cdot)\rangle$.
- To bound $R_K(\pi^*, s)$, we cast Politex as an online linear optimization for each state $s \in \mathcal{S}$:
  - In each iteration $k \in [K]$, Politex chooses a distribution $\pi_k(\cdot|s) \in \Delta_A$ for each state s.
  - The "environment" chooses and reveals the vector $\hat{q}_k(s,\cdot) \in \mathbb{R}^A$ and Politex receives a reward $\langle \pi_k(\cdot|s), \hat{q}_k(s,\cdot)\rangle$.
  - The aim is to do as well as the optimal policy π^* that receives a reward $\langle \pi^*(\cdot|s), \hat{q}_k(s,\cdot)\rangle$

Recap

Generic online optimization

- In iteration k, the algorithm chooses $w_k \in \mathcal{W}$. The environment then chooses and reveals the function $f_k : \mathcal{W} \to \mathbb{R}$ and the algorithm receives a reward $f_k(w_k)$.
- **Regret**: $R_K(w^*) := \sum_{k=0}^{K-1} [f_k(w^*) - f_k(w_k)]$.
- **Online Gradient Ascent**: $w_{k+1} = \arg\max_{w \in \mathcal{W}} \left[\langle \nabla f_k(w_k), w \rangle - \frac{1}{2\eta_k} \|w - w_k\|_2^2 \right]$.

why and what ?

- **Online Mirror Ascent**: $w_{k+1} = \arg\max_{w \in \mathcal{W}} \left[\langle \nabla f_k(w_k), w \rangle - \frac{1}{\eta_k} D_\psi(w, w_k) \right]$. Here ψ is the mirror map and $D_\psi(y, x) := \psi(y) - \psi(x) - \langle \nabla \psi(x), y - x \rangle$ is the Bregman divergence.
- Online Mirror Ascent is equivalent to the following update:
  $w_{k+1/2} = (\nabla \psi)^{-1} (\nabla \psi(w_k) + \eta_k \nabla f_k(w_k))$, $w_{k+1} = \arg\min_{w \in \mathcal{W}} D_\psi(w, w_{k+1/2})$.
- **Lipschitz continuous functions**: For all w, $\|\nabla f(w)\|_\infty \le G$
- **Strongly-convex functions**: For all y, x, $f(y) \ge f(x) + \langle \nabla f(x), y - x \rangle + \frac{\nu}{2} \|y - x\|_1^2$

Digression – Online Optimization

Claim: For G-Lipschitz linear functions $\{f_k\}_{k=0}^{K-1}$ such that $f_k(w) = \langle g_k, w \rangle$, online mirror ascent with a ν strongly-convex mirror map ψ, $\eta_k = \eta = \sqrt{\frac{2\nu}{K}} \frac{D}{G}$ where $D^2 := \max_{u \in \mathcal{W}} D_\psi(u, w_0)$ has the following regret for all $u \in \mathcal{W}$,

$$R_K(u) \leq \frac{\sqrt{2}\, DG}{\sqrt{\nu}} \sqrt{K}\,,$$

Proof: Recall the mirror ascent update: $\nabla\phi(w_{k+1/2}) = \nabla\phi(w_k) + \eta_k \nabla f_k(w_k)$.

Setting $\eta_k = \eta$ and using the definition of regret

$= \langle g_k, u - w_k \rangle$, $g_k = \nabla f_k(w_k)$; $\nabla\psi(w_{k+\frac{1}{2}}) := \nabla\psi(w_k) + \eta g_k$

$$R_K(u) = \textstyle\sum_{k=0}^{K-1} [\langle g_k, u \rangle - \langle g_k, w_k \rangle] = \sum_{k=0}^{K-1} \frac{1}{\eta} \left\langle \nabla\psi(\underbrace{w_{k+1/2}}_{y}) - \nabla\psi(\underbrace{w_k}_{z}), \underbrace{u}_{x} - w_k \right\rangle.$$

Using the three point Bregman property: for any 3 points x, y, z,
$\langle \nabla\psi(z) - \nabla\psi(y), z - x \rangle = D_\psi(x, z) + D_\psi(z, y) - D_\psi(x, y)$,

$$\left\langle \nabla\psi(w_{k+1/2}) - \nabla\psi(w_k), u - w_k \right\rangle = D_\psi(u, w_k) + D_\psi(w_k, w_{k+1/2}) - D_\psi(u, w_{k+1/2})$$

$$\implies R_K(u) = \sum_{k=0}^{K-1} \frac{1}{\eta} \left[D_\psi(u, w_k) + D_\psi(w_k, w_{k+1/2}) - D_\psi(u, w_{k+1/2}) \right]$$

Digression – Online Optimization

$R_K(u) = \sum_{k=0}^{K-1} \frac{1}{\eta} \left[D_\psi(u, w_k) + D_\psi(w_k, w_{k+1/2}) - D_\psi(u, w_{k+1/2})\right]$, $w_{k+1} = \arg\min_{w \in \mathcal{W}} D_\psi(w, w_{k+1/2})$.

Recall the optimality condition: for convex f, if $x^* = \arg\min_{x \in \mathcal{X}} f(x)$, then $\forall x \in \mathcal{X}$, $\langle \nabla f(x^*), x^* - x \rangle \leq 0$. Q: Why is $D_\psi(w, w_{k+1/2})$ convex in w? 见笔记 Using the above condition for $f = D_\psi(w, w_{k+1/2})$ and $x^* = w_{k+1}$, we infer that for any $w \in \mathcal{W}$,

$$\left\langle \nabla\psi(\underset{z}{w_{k+1}}) - \nabla\psi(\underset{y}{w_{k+1/2}}), \underset{z}{w_{k+1}} - \underset{x}{w} \right\rangle \leq 0$$

$$\implies D_\psi(w, w_{k+1}) + D_\psi(w_{k+1}, w_{k+1/2}) - D_\psi(w, w_{k+1/2}) \leq 0 \qquad \text{(3 point Bregman property)}$$

$$\implies -D_\psi(u, w_{k+1/2}) \leq -D_\psi(u, w_{k+1}) - D_\psi(w_{k+1}, w_{k+1/2}) \qquad \text{(Setting } w = u\text{)}$$

Putting everything together,

$$R_K(u) \leq \sum_{k=0}^{K-1} \frac{1}{\eta} \left[D_\psi(u, w_k) - D_\psi(u, w_{k+1})\right] + \left[D_\psi(w_k, w_{k+1/2}) - D_\psi(w_{k+1}, w_{k+1/2})\right]$$

$$\leq \frac{1}{\eta} D_\psi(u, w_0) + \frac{1}{\eta} \sum_{k=0}^{K-1} \left[D_\psi(w_k, w_{k+1/2}) - D_\psi(w_{k+1}, w_{k+1/2})\right]$$

Digression – Online Optimization

Recall that $R_K(u) \leq \frac{1}{\eta} D_\psi(u, w_0) + \frac{1}{\eta} \sum_{k=0}^{K-1} \left[D_\psi(w_k, w_{k+1/2}) - D_\psi(w_{k+1}, w_{k+1/2}) \right]$. By def. of D_ψ,

def   strong convexity of ψ $\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y-x \rangle + \frac{\nu}{2} \|y-x\|_1^2$

$$D_\psi(w_k, w_{k+1/2}) - D_\psi(w_{k+1}, w_{k+1/2}) = \psi(w_k) - \psi(w_{k+1}) - \langle \nabla\psi(w_{k+1/2}), w_k - w_{k+1} \rangle$$

偷带笔迹的课件吧……

$$\leq \langle \nabla\psi(w_k) - \nabla\psi(w_{k+1/2}), w_k - w_{k+1} \rangle - \frac{\nu}{2} \|w_k - w_{k+1}\|_1^2$$

(Using strong-convexity of ψ with $y = w_{k+1}$ and $x = w_k$)

$\nabla\psi(w_{k+\frac{1}{2}}) = \nabla\psi(w_k) + \eta \cdot g_k$

$$= -\eta \langle g_k, w_k - w_{k+1} \rangle - \frac{\nu}{2} \|w_k - w_{k+1}\|_1^2 \quad \text{(Using the mirror ascent update)}$$

$\leq -\eta [\langle g_k, w_k - w_{k+1} \rangle]$

$$\leq \eta G \|w_k - w_{k+1}\|_1 - \frac{\nu}{2} \|w_k - w_{k+1}\|_1^2$$

(Holder's inequality: $\langle x, y \rangle \leq \|x\|_\infty \|y\|_1$ and since f_k is G-Lipschitz)

$$\leq \frac{\eta^2 G^2}{2\nu} \quad \text{(For all } z\text{, } az - bz^2 \leq \tfrac{a^2}{4b}\text{)}$$

$$\implies R_K(u) \leq \frac{1}{\eta} D_\psi(u, w_0) + \frac{\eta G^2 K}{2\nu} \leq \frac{D^2}{\eta} + \frac{\eta G^2 K}{2\nu} \quad \text{(Since } D_\psi(u, w_0) \leq D^2\text{)}$$

$$R_K(u) \leq \frac{\sqrt{2} DG}{\sqrt{\nu}} \sqrt{K} \quad \square \qquad \text{(Setting } \eta = \sqrt{\tfrac{2\nu}{K}} \tfrac{D}{G}\text{)}$$

Convergence of Politex

- We have proved that: For G-Lipschitz linear functions $\{f_k\}_{k=0}^{K-1}$ such that $f_k(w) = \langle g_k, w \rangle$, online mirror ascent with a ν strongly-convex mirror map ψ, $\eta_k = \eta = \sqrt{\frac{2\nu}{K}}\, \frac{D}{G}$ where $D^2 := \max_{u \in \mathcal{W}} D_\psi(u, w_0)$ has the following regret for all $u \in \mathcal{W}$, $R_K(u) \leq \frac{\sqrt{2}DG}{\sqrt{\nu}}\sqrt{K}$.
- For Politex (for $s \in \mathcal{S}$), $w = \pi_s := \pi(\cdot|s)$, $\mathcal{W} = \Delta_A$, $g_k = \hat{q}_k(s, \cdot)$ and $u = \pi^*_s := \pi^*(\cdot|s)$.

Claim 1: For policies $\pi, \tilde{\pi}$, if $\pi_s := \pi(\cdot|s) \in \Delta_A$, with the *negative entropy mirror map* equal to: $\psi(\pi_s) = \sum_{a \in \mathcal{A}} \pi(a|s) \log(\pi(a|s))$, the corresponding Bregman divergence $D_\psi(\pi_s, \tilde{\pi}_s)$ is equal to the KL divergence equal to: $\text{KL}(\pi_s || \tilde{\pi}_s) = \sum_{a \in \mathcal{A}} \pi(a|s) \log\left(\pi(a|s) / \tilde{\pi}(a|s)\right)$..

Claim 2: For an arbitrary state $s \in \mathcal{S}$, prove that at iteration $k \geq 0$, online mirror ascent with $w = \pi(\cdot|s) \in \mathbb{R}^A$, negative entropy mirror map, step-size $\eta_k = \eta$ for all k has the following *multiplicative weights* update on linear losses $f_k(\pi(\cdot|s)) = \langle \pi(\cdot|s), \hat{q}_k(s, \cdot) \rangle$ for all $a \in \mathcal{A}$, $\pi_{k+1}(a|s) = \frac{\pi_k(a|s)\, \exp(\eta\, \hat{q}_k(s,a))}{\sum_{a' \in \mathcal{A}} \pi_k(a'|s)\, \exp(\eta\, \hat{q}_k(s,a'))}$

Claim 3: With $\pi_0(a|s) = \frac{1}{A}$ for each (s, a), the above update is equal to the update for Politex.

Prove in Assignment 3!

Convergence of Politex

Using the claims on the previous slide, we can conclude that Politex (for state $s \in \mathcal{S}$) has the following regret: $R_K(\pi_s^*) \leq \frac{\sqrt{2}DG}{\sqrt{\nu}} \sqrt{K}$. We now need to characterize the constants D, G, ν.

• Recall that $D^2 = \max D_\psi(u, w_0) = \mathrm{KL}(\pi^*(\cdot|s) || \pi_0(\cdot|s))$. For all $a \in \mathcal{A}$, choose $\pi_0(a|s) = \frac{1}{A}$ i.e. for each state, π_0 is a uniform distribution over actions. With this choice,

$$\mathrm{KL}(\pi^*(\cdot|s)||\pi_0(\cdot|s)) = \sum_a \pi^*(a|s) \log\left(A\,\pi^*(a|s)\right) \leq \log\left(A \max_a \pi^*(a|s)\right) \sum_a \pi^*(a|s) \leq \log(A)$$

• Recall that $\|\nabla f(x)\|_\infty \leq G$. If the $\hat{q}_k(s,a)$ functions are constrained to lie in the $[0, 1/1-\gamma]$ interval, then $G = \frac{1}{1-\gamma}$.

• Recall that ν is the strong-convexity of ψ, i.e. the following inequality holds: $\psi(y) \geq \psi(x) + \langle \nabla\psi(x), y - x\rangle + \frac{\nu}{2}\,\|y - x\|_1^2$.

$$\psi(y) - \psi(x) - \langle \nabla\psi(x), y - x\rangle = D_\psi(y, x) = \mathrm{KL}(y||x) \geq \frac{1}{2}\,\|y - x\|_1^2 \quad \text{(Pinsker's inequality)}$$

Hence, $\nu = 1$.

Convergence of Politex

Putting everything together, we can prove the following claim:

Claim: If $\hat{q}(s,a) \in [0, 1/1-\gamma]$ for all (s,a), Politex with $\pi_0(a|s) = \frac{1}{A}$ for all (s,a) and $\eta_k = \eta = \sqrt{\frac{2\log(A)}{K}}(1-\gamma)$ has the following regret,

$$R_K(\pi^*, s) \leq \frac{\sqrt{2\log(A)}}{1-\gamma}\sqrt{K} \implies \|\text{Regret}(K)\|_\infty = \frac{\sqrt{2\log(A)}}{1-\gamma}\sqrt{K}$$

Combining the above bound with the general result for Politex,

$$\left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{\sqrt{2\log(A)}}{(1-\gamma)^2\sqrt{K}} + \frac{2\max_{k\in\{0,\dots,K-1\}}\|\epsilon_k\|_\infty}{(1-\gamma)}$$

Controlling the policy evaluation error using G experimental design and Monte-Carlo estimation ensures that $\max_{k\in\{0,\dots,K-1\}}\|\epsilon_k\|_\infty \leq \varepsilon_{\mathbf{b}}\left(1+\sqrt{d}\right) + \varepsilon_{\mathbf{s}}\sqrt{d}$.

$$\implies \left\|v^{\bar{\pi}_K} - v^*\right\|_\infty \leq \frac{\sqrt{2\log(A)}}{(1-\gamma)^2\sqrt{K}} + \frac{2\varepsilon_{\mathbf{b}}\left(1+\sqrt{d}\right) + 2\varepsilon_{\mathbf{s}}\sqrt{d}}{(1-\gamma)}$$

Policy Gradient

Policy Gradient

- For approximate policy iteration and Politex, we parameterized the q functions, and designed algorithms that avoid the explicit dependence on S.

- Policy gradient methods directly parameterize the policy and use gradient ascent to maximize the value function. Formally, given a policy parameterization s.t. $\pi = h(\theta)$ and a step-size η, policy gradient methods have the following update:

$$\theta_{t+1} = \theta_t + \eta \nabla_\theta J(\theta_t) \quad \text{where} \quad J(\theta) := v^{\pi_\theta}(\rho) = \mathbb{E}_{s_0 \sim \rho} v^{\pi_\theta}(s_0)$$

- Common policy parameterizations include:
  - **Tabular softmax policy parameterization**: $\forall (s,a) \in \mathcal{S} \times \mathcal{A}$, there is a parameter $\theta(s,a)$ s.t. $\pi(a|s) = \frac{\exp(\theta(s,a))}{\sum_{a'} \exp(\theta(s,a'))}$
  - **Log-linear policies**: Given access to features $\Phi \in \mathbb{R}^{SA \times d}$, $\pi(a|s) = \frac{\exp(\langle \phi(s,a), \theta \rangle)}{\sum_{a'} \exp(\langle \phi(s,a'), \theta \rangle)}$ for parameter $\theta \in \mathbb{R}^d$.
  - **Energy-based policies**: Using a general function approximation (deep neural network) $f_\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$, $\pi(a|s) = \frac{\exp(f_\theta(s,a))}{\sum_{a'} \exp(f_\theta(s,a')))}$.

Policy Gradient

In order to calculate $\nabla J(\theta)$ for a general policy parameterization, we recall the definitions of the *state occupancy measure* $d^\pi \in \mathbb{R}^S$ and the *state-action occupancy measure* $\mu^\pi \in \mathbb{R}^{S \times A}$.

$$\mu^\pi(s,a) := (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{t=0}^{\infty} \gamma^t \, \Pr[S_t = s, A_t = a | S_0 = s_0]$$

$$d^\pi(s) := (1-\gamma) \sum_{s_0 \in \mathcal{S}} \rho(s_0) \sum_{t=0}^{\infty} \gamma^t \, \Pr[S_t = s | S_0 = s_0]$$

In Assignment 2, we proved that if $r \in \mathbb{R}^{S \times A}$ is the reward vector,
(i) $v^\pi(\rho) = \frac{1}{1-\gamma} \langle \mu^\pi, r \rangle$, (ii) $d^\pi(s) = \sum_a \mu^\pi(s,a)$, (iii) $\pi(a|s) = \frac{\mu^\pi(s,a)}{\sum_{a'} \mu^\pi(s,a')}$. Hence,

$$v^\pi(\rho) = \frac{1}{1-\gamma} \sum_s d^\pi(s) \sum_a \pi(a|s) \, r(s,a) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^\pi} \mathbb{E}_{a \sim \pi(\cdot|s)} \, r(s,a)$$

Recall that $v^\pi(\rho)$ can be (approximately) computed by rolling out trajectories and using Monte-Carlo estimation. By the above equivalence, the expectation $\mathbb{E}_{s \sim d^\pi} \mathbb{E}_{a \sim \pi(\cdot|s)}$ can also be estimated similarly.

Policy Gradient Theorem

Claim: $\nabla_\theta J(\theta) = \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$.

Proof:

$$v^{\pi_\theta}(s) = \sum_a \pi_\theta(a|s)\, q^{\pi_\theta}(s,a) \implies \frac{\partial v^{\pi_\theta}(s)}{\partial\theta} = \sum_a \left[\frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a) + \pi_\theta(a|s)\, \frac{\partial q^{\pi_\theta}(s,a)}{\partial\theta}\right]$$

$$q^{\pi_\theta}(s,a) = r(s,a) + \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\, v^{\pi_\theta}(s') \implies \frac{\partial q^{\pi_\theta}(s,a)}{\partial\theta} = \gamma \sum_{s'\in\mathcal{S}} \mathcal{P}(s'|s,a)\, \frac{\partial v^{\pi_\theta}(s')}{\partial\theta}$$

$$\implies \frac{\partial v^{\pi_\theta}(s)}{\partial\theta} = \sum_a \left[\frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right] + \gamma \sum_{s'\in\mathcal{S}} \sum_a \mathcal{P}(s'|s,a)\, \pi_\theta(a|s) \frac{\partial v^{\pi_\theta}(s')}{\partial\theta}$$

$$\frac{\partial v^{\pi_\theta}(s)}{\partial\theta} = \sum_a \left[\frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right] + \gamma \sum_{s'} \mathbf{P}_{\pi_\theta}[s,s']\, \frac{\partial v^{\pi_\theta}(s')}{\partial\theta}$$

Hence, $\frac{\partial v^{\pi_\theta}(s)}{\partial\theta}$ can be expressed in terms of $\frac{\partial v^{\pi_\theta}(s')}{\partial\theta}$. We will use this result recursively from the starting state.

Policy Gradient Theorem

Recall that $\frac{\partial v^{\pi_\theta}(s)}{\partial \theta} = \sum_a \left[\frac{\partial \pi_\theta(a|s)}{\partial \theta} q^{\pi_\theta}(s,a) \right] + \gamma \sum_{s'} \mathbf{P}_{\pi_\theta}[s,s'] \frac{\partial v^{\pi_\theta}(s')}{\partial \theta}$. Starting from state s_0,

$$\frac{\partial v^{\pi_\theta}(s_0)}{\partial \theta} = \underbrace{\sum_{a_0} \left[\frac{\partial \pi_\theta(a_0|s_0)}{\partial \theta} q^{\pi_\theta}(s_0, a_0) \right]}_{:=\omega(s_0)} + \gamma \sum_{s_1} \mathbf{P}_{\pi_\theta}[s_0, s_1] \frac{\partial v^{\pi_\theta}(s_1)}{\partial \theta}$$

$$= \omega(s_0) + \gamma \sum_{s_1} \mathbf{P}_{\pi_\theta}[s_0, s_1] \left[\sum_{a_1} \left[\frac{\partial \pi_\theta(a_1|s_1)}{\partial \theta} q^{\pi_\theta}(s_1, a_1) \right] + \gamma \sum_{s_2} \mathbf{P}_{\pi_\theta}[s_1, s_2] \frac{\partial v^{\pi_\theta}(s_2)}{\partial \theta} \right]$$

$$= \omega(s_0) + \gamma \sum_{s_1} \mathbf{P}_{\pi_\theta}[s_0, s_1] \omega(s_1) + \gamma^2 \sum_{s_1} \sum_{s_2} \mathbf{P}_{\pi_\theta}[s_0, s_1] \mathbf{P}_{\pi_\theta}[s_1, s_2] \frac{\partial v^{\pi_\theta}(s_2)}{\partial \theta}$$

$$= \omega(s_0) + \gamma \sum_{s_1} \Pr[S_1 = s_1 | S_0 = s_0] \omega(s_1) + \gamma^2 \sum_{s_2} \Pr[S_2 = s_2 | S_0 = s_0] \frac{\partial v^{\pi_\theta}(s_2)}{\partial \theta}$$

$$\Longrightarrow \frac{\partial v^{\pi_\theta}(s_0)}{\partial \theta} = \sum_{t=0}^{\infty} \gamma^t \left[\sum_{s_t} \Pr[S_t = s_t | S_0 = s_0] \omega(s_t) \right] \qquad \text{(Recursively unrolling)}$$

Policy Gradient Theorem

Recall that $\frac{\partial v^{\pi_\theta}(s_0)}{\partial \theta} = \sum_{t=0}^{\infty} \gamma^t \left[\sum_{s_t} \Pr[S_t = s_t | S_0 = s_0]\, \omega(s_t)\right]$. Rearranging the sum,

$$\frac{\partial v^{\pi_\theta}(s_0)}{\partial \theta} = \sum_{s} \left[\sum_{t=0}^{\infty} \gamma^t \Pr[S_t = s | S_0 = s_0]\right] \omega(s)$$

$$\implies \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta} = \sum_{s_0} \rho(s_0) \frac{\partial v^{\pi_\theta}(s_0)}{\partial \theta} = \sum_{s_0} \rho(s_0) \sum_{s} \left[\sum_{t=0}^{\infty} \gamma^t \Pr[S_t = s | S_0 = s_0]\right] \omega(s)$$

$$= \sum_{s} \left[\sum_{s_0} \rho(s_0) \left[\sum_{t=0}^{\infty} \gamma^t \Pr[S_t = s | S_0 = s_0]\right]\right] \omega(s)$$

$$= \frac{1}{1-\gamma} \sum_{s} d^{\pi_\theta}(s)\, \omega(s) = \frac{1}{1-\gamma} \sum_{s} d^{\pi_\theta}(s) \sum_{a} \left[\frac{\partial \pi_\theta(a|s)}{\partial \theta} q^{\pi_\theta}(s,a)\right]$$

(By def. of $d^{\pi}(s)$)

$$\implies \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta} = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d^{\pi_\theta}} \left[\sum_{a \in \mathcal{A}} \frac{\partial \pi_\theta(a|s)}{\partial \theta} q^{\pi_\theta}(s,a)\right] \quad \square$$

Policy Gradient Theorem

In order to compute $\frac{\partial v^{\pi_\theta}(\rho)}{\partial\theta} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\left[\sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$ algorithmically,

let us simplify $\left[\sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$,

$$\left[\sum_{a\in\mathcal{A}} \frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right] = \left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\, \frac{1}{\pi_\theta(a|s)}\, \frac{\partial \pi_\theta(a|s)}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$$

$$= \left[\sum_{a\in\mathcal{A}} \pi_\theta(a|s)\, \frac{\partial \ln(\pi_\theta(a|s))}{\partial\theta}\, q^{\pi_\theta}(s,a)\right] = \mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\frac{\partial \ln(\pi_\theta(a|s))}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$$

$$\frac{\partial v^{\pi_\theta}(\rho)}{\partial\theta} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\pi_\theta}}\mathbb{E}_{a\sim\pi_\theta(\cdot|s)}\left[\frac{\partial \ln(\pi_\theta(a|s))}{\partial\theta}\, q^{\pi_\theta}(s,a)\right]$$

The term $\frac{\partial \ln(\pi_\theta(a|s))}{\partial\theta}$ is referred to as the *score function*.

As before, the $\mathbb{E}_{s\sim d^{\pi}}\mathbb{E}_{a\sim\pi(\cdot|s)}$ expectations can be computed by rolling out trajectories starting at $s_0\sim\rho$, taking actions $a_t\sim\pi_\theta(\cdot|s_t)$ for $t\geq 0$ and using Monte-Carlo estimation. The gradient expression involves $q^{\pi}(s,a)$ that can be estimated using a policy evaluation method such as TD.

Softmax Policy Gradient

The policy gradient theorem gives us a handle on $\nabla_\theta J(\theta)$ enabling us to use the resulting update.

In order to analyze the convergence of policy gradient, we will only focus on the tabular softmax policy parameterization in this course.

Tabular softmax policy parameterization: Consider $\theta \in \mathbb{R}^A$ and the function $h : \mathbb{R}^A \to \mathbb{R}^A$ such that $h(\theta) = \pi_\theta$ where $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$. For the tabular softmax policy parameterization, $\pi_\theta(\cdot|s) = h(\theta(s, \cdot))$.

Claim: The Jacobian of $h : \mathbb{R}^A \to \mathbb{R}^A$ is given by $H(\pi_\theta) \in \mathbb{R}^{A \times A} = \text{diag}(\pi_\theta) - \pi_\theta \, \pi_\theta^T$ where $\text{diag}(\pi_\theta) \in \mathbb{R}^{A \times A}$ is a diagonal matrix s.t. $[\text{diag}(\pi_\theta)]_{a,a} = \pi_\theta(a)$ and $\pi_\theta \in \mathbb{R}^A$ s.t. $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$.

Prove in Assignment 4!

Let us first instantiate the policy gradient expression with this choice of the policy parameterization.

Softmax Policy Gradient

Claim: For the tabular softmax policy parameterization,

$$\frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{d^{\pi_\theta}(s)}{1-\gamma}\, \pi_\theta(a|s)\, \mathfrak{a}^{\pi_\theta}(s,a)\,,$$

where $\mathfrak{a}^{\pi_\theta}(s,a) = q^{\pi_\theta}(s,a) - v^{\pi_\theta}(s)$ is the advantage (over π_θ) of taking action a in state s.

Proof: For vector θ, we know that $\frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta} = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d^{\pi_\theta}} \left[\sum_{a' \in \mathcal{A}} \frac{\partial \pi_\theta(a'|s')}{\partial \theta}\, q^{\pi_\theta}(s',a')\right]$.

For the tabular softmax policy parameterization, $H(\pi_\theta) = \frac{\partial \pi_\theta}{\partial \theta} = \mathrm{diag}(\pi_\theta) - \pi_\theta\, \pi_\theta^T$.

Since there is no coupling between the parameters $\theta(s,a)$, for $s' \neq s$ and any $a \in \mathcal{A}$, $\pi_\theta(a|s')$ does not depend on $\theta(s,a)$ and hence, $\frac{\partial \pi_\theta(a|s'))}{\partial \theta(s,\cdot)} = \mathbf{0}$.

$$\frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,\cdot)} = \frac{d^{\pi_\theta}(s)}{1-\gamma} \sum_{a' \in \mathcal{A}} \frac{\partial \pi_\theta(a'|s)}{\partial \theta(s,\cdot)}\, q^{\pi_\theta}(s,a') = \frac{d^{\pi_\theta}(s)}{1-\gamma} \underbrace{\frac{\partial \pi_\theta(\cdot|s)}{\partial \theta(s,\cdot)}}_{A \times A} \underbrace{q^{\pi_\theta}(s,\cdot)}_{A \times 1}$$

$$= \frac{d^{\pi_\theta}(s)}{1-\gamma} H(\pi_\theta(\cdot|s))\, q^{\pi_\theta}(s,\cdot) = \frac{d^{\pi_\theta}(s)}{1-\gamma} \left[\mathrm{diag}(\pi_\theta(\cdot|s)) - \pi_\theta(\cdot|s)\, \pi_\theta(\cdot|s)^T\right] q^{\pi_\theta}(s,\cdot)$$

Softmax Policy Gradient

Recall that $\frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,\cdot)} = \frac{d^{\pi_\theta}(s)}{1-\gamma} \left[\text{diag}(\pi_\theta(\cdot|s)) - \pi_\theta(\cdot|s)\pi_\theta(\cdot|s)^T\right] q^{\pi_\theta}(s,\cdot)$. Define $\omega \in \mathbb{R}^A := \left[\pi_\theta(a_1|s)\, q^{\pi_\theta}(s,a_1), \pi_\theta(a_2|s)\, q^{\pi_\theta}(s,a_2) \ldots \pi_\theta(a_A|s)\, q^{\pi_\theta}(s,a_A)\right]$. Hence,

$$\frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,\cdot)} = \frac{d^{\pi_\theta}(s)}{1-\gamma} \left[\omega - \left[\sum_{a'} \pi_\theta(a'|s)\, q^{\pi}(s,a')\right] \pi_\theta(\cdot|s)\right]$$

Taking the component corresponding to action a,

$$\implies \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(s,a)} = \frac{d^{\pi_\theta}(s)}{1-\gamma} \left[\pi_\theta(a|s)\, q^{\pi_\theta}(s,a) - \pi_\theta(a|s)\, v^{\pi_\theta}(s)\right]$$

$$= \frac{d^{\pi_\theta}(s)}{1-\gamma} \pi_\theta(a|s)\, \mathfrak{a}^{\pi}_{\theta}(s,a) \quad \square$$

Softmax Policy Gradient for Bandits

In order to analyze the convergence of softmax policy gradient, let us further simplify the problem and focus on the special case of multi-armed bandits where $\gamma = 0$ and $S = 1$. In this case, assuming that the rewards $r \in \mathbb{R}^A$ are deterministic,

$$J(\theta) = \mathbb{E}_{a \sim \pi_\theta}[r(a)] = \langle \pi_\theta, r \rangle$$

For the tabular softmax parameterization, $\theta \in \mathbb{R}^A$ and $\pi_\theta = h(\theta)$. In this case, $q^{\pi_\theta} \in \mathbb{R}^A = r$ and $\mathfrak{a}^{\pi_\theta} \in R^A = r - \langle \pi_\theta, r \rangle$. Hence,

$$\frac{\partial J(\theta)}{\partial \theta(a)} = \frac{\partial v^{\pi_\theta}(\rho)}{\partial \theta(a)} = \pi_\theta(a) \left[r(a) - \langle \pi_\theta, r \rangle \right]$$

Hence, for multi-armed bandit problems, the softmax policy gradient with a tabular parameterization can be written as: $\theta_{t+1} = \theta_t + \eta \left[\pi_\theta(a) \left[r(a) - \langle \pi_\theta, r \rangle \right] \right]$.

bandit + softmax policy gradient + tabular parameterization + deterministic

Q: Why is this algorithm impractical from a bandits perspective?

Next, we will see that even for this special case, $J(\theta)$ is non-concave in θ. This implies that in general, $J(\theta)$ is a non-concave function of θ when using the softmax parameterization.

Softmax Policy Gradient for Bandits

Claim: For the tabular softmax policy parameterization where $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$, the objective $J(\theta) = \langle \pi_\theta, r \rangle$ can be non-concave w.r.t θ.

Proof: Recall that a function $f : \mathcal{D} \to \mathbb{R}$ is concave if for all $\theta, \theta' \in \mathcal{D}$ and $\alpha \in [0, 1]$, $f(\alpha\theta + (1-\alpha)\theta') \geq \alpha f(\theta) + (1-\alpha) f(\theta')$. Consider a multi-armed bandit problem where $A = 3$, and $r = [1, 9/10, 1/10]$, $\theta = [0, 0, 0]$ and $\theta' = [\ln(9), \ln(16), \ln(25)]$. Choosing $\alpha = \frac{1}{2}$,

$$\pi = h(\theta) = [1/3, 1/3, 1/3] \implies J(\theta) = \frac{1}{3} + \frac{3}{10} + \frac{1}{30} = \frac{2}{3}$$

$$\pi' = h(\theta') = [9/50, 16/50, 25/50] \implies J(\theta) = \frac{90}{500} + \frac{144}{500} + \frac{25}{500} = \frac{259}{500}$$

$$\implies \text{RHS} = \alpha J(\theta) + (1-\alpha) J(\theta') = \frac{1}{2}\left(\frac{2}{3} + \frac{259}{500}\right) = \frac{1777}{3000}$$

$$\alpha\theta + (1-\alpha)\theta' = [\ln(3), \ln(4), \ln(5)] \implies h(\alpha\theta + (1-\alpha)\theta') = [3/12, 4/12, 5/12]$$

$$\implies \text{LHS} = J(\alpha\theta + (1-\alpha)\theta') = \frac{3}{12} + \frac{36}{120} + \frac{5}{120} = \frac{71}{120}.$$

RHS $= \frac{1777}{3000} = \frac{14216}{24000} > \frac{14200}{24000} =$ LHS, meaning that $J(\theta)$ is non-concave for this example.

Digression – Smooth functions

Smooth functions: For smooth functions that are differentiable everywhere, the gradient is Lipschitz-continuous i.e. it can not change arbitrarily fast.

- Formally, the gradient ∇f is L-Lipschitz continuous if for all $x, y \in \mathcal{D}$,

$$\|\nabla f(x) - \nabla f(y)\| \leq L \, \|x - y\|$$

where L is the Lipschitz constant of the gradient (also called the smoothness constant of f).

- If f is twice-differentiable and smooth, then for all $x \in \mathcal{D}$, $\nabla^2 f(x) \preceq L \, I_d$ i.e. $\sigma_{\max}[\nabla^2 f(x)] \leq L$ where $\sigma_{\max}$ is the maximum singular value.

- For L-smooth functions, for all $x, y \in \mathcal{D}$,

$$|f(y) - f(x) - \langle \nabla f(x), y - x\rangle| \leq \frac{L}{2} \|y - x\|^2$$

Hence the function $f(y)$ is upper and lower-bounded by quadratics: $f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2} \|y - x\|^2$ and $f(x) + \langle \nabla f(x), y - x\rangle - \frac{L}{2} \|y - x\|^2$ respectively. These bounds are *global* and hold for all $y \in \mathcal{D}$.

Softmax Policy Gradient

Fact: For the tabular softmax policy parameterization where $\pi_\theta = h(\theta)$ i.e. $\pi_\theta(a) = \frac{\exp(\theta(a))}{\sum_{a'} \exp(\theta(a'))}$, the objective $J(\theta) = \langle \pi_\theta, r \rangle$ is $\frac{5}{2}$-smooth.

See [MXSS20, Lemmas 2] for a proof. Such a smoothness property also holds for general MDPs (see [MXSS20, Lemma 7]).

• By putting together these results, we conclude that for the tabular softmax policy parameterization, the objective $J(\theta)$ is a smooth, non-concave function.

• Hence, in general (without additional properties), policy gradient is not guaranteed to converge to the optimal policy, but only to a stationary point where $\|\nabla_\theta J(\theta)\| = 0$. Assuming that we can exactly calculate $\nabla_\theta J(\theta)$, we can prove the following standard result from non-convex optimization.

Claim: For the tabular softmax policy parameterization where $J(\theta)$ is L-smooth w.r.t θ, softmax policy gradient with $\eta = \frac{1}{L}$ returns $\hat{\theta}_T$ such that $\left\| \nabla J(\hat{\theta}_T) \right\|^2 \leq \epsilon$ and requires $T = \frac{2L}{(1-\gamma)\,\epsilon}$ iterations.

Stationary point Convergence of Softmax Policy Gradient

Proof: Using the L-smoothness of J with $x = \theta_t$ and $y = \theta_{t+1} = \theta_t + \frac{1}{L}\nabla J(\theta_t)$ in the quadratic bound (also referred to as the *ascent lemma*),

$$J(\theta_{t+1}) \geq J(\theta_t) + \left\langle \nabla J(\theta_t), \frac{1}{L}\nabla J(\theta_t) \right\rangle - \frac{L}{2}\left\| \frac{1}{L}\nabla J(\theta_t) \right\|^2$$
$$\implies J(\theta_{t+1}) \geq J(\theta_t) + \frac{1}{2L}\left\| \nabla J(\theta_t) \right\|^2$$

By moving from θ_t to θ_{t+1}, the algorithm has increased the value of J. Rearranging the inequality, for every iteration t,

$$\frac{1}{2L}\left\| \nabla J(\theta_t) \right\|^2 \leq J(\theta_{t+1}) - J(\theta_t)$$

Summing up from $t = 0$ to $T - 1$,

$$\frac{1}{2L}\sum_{t=0}^{T-1}\left\| \nabla J(\theta_t) \right\|^2 \leq \sum_{t=0}^{T-1}[J(\theta_{t+1}) - J(\theta_t)] = J(\theta_T) - J(\theta_0)$$

Stationary point Convergence of Softmax Policy Gradient

Recall that $\frac{1}{2L}\sum_{t=0}^{T-1}\left\|\nabla J(\theta_t)\right\|^2 \leq J(\theta_T) - J(\theta_0)$. Since $J(\theta) \in \left[0, \frac{1}{1-\gamma}\right]$ for all θ,

$$\frac{\sum_{t=0}^{T-1}\left\|\nabla J(\theta_t)\right\|^2}{T} \leq \frac{2L}{(1-\gamma)\,T}$$

Define $\hat{\theta}_T := \arg\min_{t\in\{0,1,\ldots,T-1\}}\left\|\nabla J(\theta_t)\right\|^2$.

$$\left\|\nabla J(\hat{\theta}_T)\right\|^2 \leq \frac{2L}{(1-\gamma)\,T}$$

If the RHS equal to $\frac{2L}{(1-\gamma)\,T} \leq \epsilon$, this would guarantee that $\left\|\nabla J(\hat{\theta}_T)\right\|^2 \leq \epsilon$ and we would achieve our objective. Hence, we need to run the algorithm for $T \geq \frac{2L}{(1-\gamma)\,\epsilon}$ iterations.

Next, we will see that for the tabular softmax policy parameterization, the objective $J(\theta)$ satisfies an additional non-uniform gradient domination property that allows us to prove convergence to the optimal policy.

References i

Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans, *On the global convergence rates of softmax policy gradient methods*, International Conference on Machine Learning, PMLR, 2020, pp. 6820–6829.