March 11

Gradient Computation

PG - Theorem

Bhandari - Russo

Silver

Sutton / Salvador / Kakade / ...

Deterministic policies $\pi_\theta : S \to \mathcal{A}$

$\mathcal{A} \subseteq \mathbb{R}^p$

Finite-action MDPs $\tilde{\mathcal{A}} = \mathcal{M}_1(\mathcal{A}) = \{p \in [0,1]^{\mathcal{A}} \mid \sum_{a=1}^{A} p_a = 1\}$

$M = (S, \mathcal{A}, P, r)$

(P, r) $P = (P_a(s))_{a \in \mathcal{A}}$

$\pi : S \to \mathcal{M}_1(\mathcal{A})$

$\pi : S \to \tilde{\mathcal{A}}$

$\tilde{P} = (\tilde{P}_{\tilde{a}})_{\tilde{a} \in \tilde{\mathcal{A}}}$ $\quad \tilde{P}_{\tilde{a}}(s,s') = \sum_{i=1}^{A} \tilde{a}_i P_i(s,s')$

$\tilde{r} = (r_{\tilde{a}})_{\tilde{a} \in \tilde{\mathcal{A}}}$ $\quad r_{\tilde{a}}(s) = \sum_{i=1}^{A} \tilde{a}_i r_i(s)$

$\tilde{M} = (S, \tilde{\mathcal{A}}, \tilde{P}, \tilde{r})$

π SML of M $\to \tilde{\pi}$ DML of $\tilde{M}$

$\tilde{\pi}(s) = \pi(s)$

$v^\pi = \tilde{v}^\pi \qquad v^* = \tilde{v}^* \qquad q^\pi = \tilde{q}^\pi$

det. policies of $\tilde{M}$ suffice for M.

$A \subseteq \mathbb{R}^p$, $\pi_\theta : S \to A$, $\theta \in \mathbb{R}^d$, $x \in \mathbb{R}^d$

Theorem: $f_\pi : x \mapsto \tilde{D}_\mu^{\pi} M_{\pi_x} q^\pi$

$\mathbb{R} \quad \supset \quad []$

$g_\pi : x \mapsto \tilde{D}_\mu^{\pi_x} v^\pi$

$$(M_\pi q)(s) = q(s, \pi(s))$$

Fix $\theta_0 \in \mathbb{R}^d$, $J(\theta) = \mu v^{\pi_\theta}$

Assume $f_{\pi_{\theta_0}}, g_{\pi_{\theta_0}} \in C^1$ @ θ_0

$$\Rightarrow \nabla J(\theta_0) = \frac{d}{dx} \tilde{D}_\mu^{\pi_{\theta_0}} M_{\pi_x} q^{\pi_{\theta_0}} \Big|_{x=\theta_0}$$

$$= \frac{d}{dx} \sum_s \tilde{D}_\mu^{\pi_\theta}(s)\, q^{\pi_{\theta_0}}(s, \pi_x(s)) \Big|_{x=\theta_0}$$

$f(u, v)$ $\quad \frac{d}{dx} f(x, x) = \frac{\partial}{\partial u} f(x, x) + \frac{\partial}{\partial v} f(x, x)$

Proof: $v^{\pi'} - v^\pi = (I - \gamma P_{\pi'})^{-1} r_{\pi'} - v^\pi$

$= (I - \gamma P_{\pi'})^{-1} \left[r_{\pi'} - (I - \gamma P_{\pi'}) v^\pi \right]$

$= (I - \gamma P_{\pi'})^{-1} \left[T_{\pi'} v^\pi - v^\pi \right]$

$\pi' = \pi_x$, $\pi = \pi_\theta$

$T_{\pi'} v^\pi = M_{\pi'}(r + \gamma P v^\pi)$ $\qquad M_{\pi'} q^\pi$

q^π

$$\mu\left(v^{\pi_x} - v^{\pi_\theta}\right) = \tilde{\nu}_\mu^{\pi_x}\left[M_{\pi_x} q^{\pi_\theta} - v^{\pi_\theta}\right] \quad / \frac{d}{dx}$$

$$f(u,v) = \tilde{\nu}_\mu^{\pi_u} M_{\pi_v} q^{\pi_\theta}$$

$$\frac{d}{dx} f(x,x) = \frac{d}{du} \tilde{\nu}_\mu^{\pi_u} M_{\pi_x} q^{\pi_\theta}\Big|_{u=x} + \frac{d}{dv} \tilde{\nu}_\mu^{\pi_x} M_{\pi_v} q^{\pi_\theta}\Big|_{v=x}$$

$$\frac{d}{dx} J(x) = \frac{d}{du} \tilde{\nu}_\mu^{\pi_u} M_{\pi_x} q^{\pi_\theta}\Big|_{u=x} + \frac{d}{dv} \tilde{\nu}_\mu^{\pi_x} M_{\pi_v} q^{\pi_\theta}\Big|_{v=x} - \frac{d}{dx} \tilde{\nu}_\mu^{\pi_x} v^{\pi_\theta}$$

$M_{\pi_x} q^{\pi_\theta} = v^{\pi_\theta}$, $x = \theta$

$$x = \theta = \theta_0$$

$$\nabla J(\theta_0) = \frac{d}{dv} \tilde{\nu}_\mu^{\pi_{\theta_0}} M_{\pi_v} q^{\pi_{\theta_0}}\Big|_{v=\theta_0} \quad // \text{Qu.e.d.}$$

$\pi_\theta : S \to \mathcal{M}_1(A)$; $\pi_\theta(a|s)$

$\nabla J(\theta)$

$$\frac{d}{dv} \tilde{\nu}_\mu^{\pi_\theta} M_{\pi_v} q^{\pi_\theta}\Big|_{v=\theta} = \frac{d}{dv} \sum_s \tilde{\nu}_\mu^{\pi_\theta}(s) \sum_a \pi_v(a|s)\, q^{\pi_\theta}(s,a)\Big|_{v=\theta}$$

$$= \sum_s \tilde{\nu}_\mu^{\pi_\theta}(s) \sum_a \frac{d}{dv} \pi_v(a|s)\Big|_{v=\theta} q^{\pi_\theta}(s,a)$$

$\sum \pi(a|s) = 1$

$$\frac{d}{dv} \log \pi_v(a|s) = \frac{\frac{d}{dv}\pi_v(a|s)}{\pi_v(a|s)}$$

$$\nabla J(\theta) = \sum_s \tilde{\rho}_\mu^{\pi_\theta}(s) \sum_a \pi_\theta(a|s)\, q^{\pi_\theta}(s,a) \frac{\partial}{\partial v} \log \pi_v(a|s) \Big|_{v=\theta}$$

$\pi_\theta(a|s) > 0$

Simulation to get $\hat{\nabla} J(\theta)$

$$(1-\gamma)\tilde{\nu}_\mu^{\pi_\theta}(s)\pi(a|s) = (1-\gamma)\nu_\mu^{\pi}(s,a)$$

$$\sum_{s,a} = 1$$

(1) $(S_i, A_i) \sim (1-\gamma)\nu_\mu^{\pi_\theta}(\ ), \quad i = 1, \dots, n$

(2) Rollout from S_i, A_i with π_θ

for $H_{\gamma,\varepsilon(1-\gamma)}$ for ε error, $\hat{Q}_{i,j}$ $j = 1, \dots, m$ indep.

$N \sim \text{Geo}(\gamma)$

$$\widehat{\nabla J(\theta)} = \frac{1}{(1-\gamma)nm} \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{Q}_{i,j} \frac{\partial}{\partial v} \log \pi_v(A_i|S_i) \Big|_{v=\theta}$$

$\frac{1}{(1-\gamma)nm}$: $\frac{1}{1-\gamma}$

$\hat{Q}_{i,j}$: $\frac{1}{1-\gamma}$

Vanilla PG : too bad. / Frank-Wolfe

NPG : Natural Policy Gradient

$$F(\theta) = \nu_\mu^{\pi_\theta} \left(\nabla_\theta \log \pi_\theta(\cdot|\cdot) \; \nabla_\theta \log \pi_\theta(\cdot|\cdot)^T \right) \in \mathbb{R}^{d \times d}$$

NPG: $\theta_{t+1} = \theta_t + \eta_t \, F(\theta_t)^\dagger \nabla_\theta J(\theta_t)$

Thm: $(1-\gamma) F(\theta)^\dagger \nabla_\theta J(\theta) =$

$$= \underset{\substack{w \in \mathbb{R}^d \\ \min \text{ norm } (\|\cdot\|_2)}}{\operatorname{argmin}} \; \nu_\mu^{\pi_\theta} \left(w^T \underbrace{\nabla_\theta \log \pi_\theta(a|s)}_{\in \mathbb{R}^d} - \widetilde{a}^{\pi_\theta} \right)^2$$

$$a^{\pi_\theta} = q^{\pi_\theta} - v^{\pi_\theta}$$

✓

Proof:

$$\pi_\theta(a|s) \propto \exp(\theta^T \varphi(s,a))$$

$$\nabla \log \pi_\theta(a|s) = \underbrace{\varphi(s,a) - \sum_{a'} \pi_\theta(a|s)\varphi(s,a)}_{\psi_\theta(s,a)}$$

NPG: $\Delta\theta_t = \eta w_t$ $\quad w_t = \arg\min_w \nu_\mu^\pi [w^T \psi_\theta - A^{\pi_\theta}]^2$

↓

Tabular; $\varphi \neq$ tabular

Q-NPG: $\Delta\theta_t = \eta w_t$ $\quad w_t = \arg\min_w \nu_\mu^\pi [w^T \varphi - Q^{\pi_\theta}]^2$

↓

Politex $\Delta\theta_t = \eta w_t$ $\quad w_t = \arg\min_w \tilde{\nu}^\pi [w^T \varphi - Q^{\pi_\theta}]^2$

$\left\| \frac{d\nu_\mu^{\pi^*}}{d\mu} \right\|_\infty = C$

Mirror-descent

$\frac{\sqrt{d\varepsilon}}{1-\gamma}$

$\frac{C\varepsilon}{(1-\gamma)^2\delta}$

Questions: ① What parametrization? Escort map?

$s \in \mathbb{R}^A$

$\frac{|s_i|^{1/2}}{\sum_{j=1}^{d} |s_j|^{1/2}}$

② Adaptive directions/stepsizes