Policy Gradient Methods and Non-Convex Optimization

Presented by :
Atmani Hanan

University mohammed vi polytechnic

July 9

Presentation plan

General introduction and motivation

- Pour une distribution ρ définie sur les états, nous définissons :

$$V^{\pi}(\rho) := \mathbb{E}_{s_0 \sim \rho}\left[V^{\pi}(s_0)\right]$$

  Considérons une classe de politiques $\left\{\pi_\theta \mid \theta \in \Theta \subset \mathbb{R}^d\right\}$, où $d = |A||S|$.
  Le problème d'optimisation sur lequel nous nous concentrons maintenant s'écrit :

$$\max_{\theta \in \Theta} V^{\pi_\theta}(\rho) \qquad \text{(P1)}$$

- Une politique déterministe π_θ n'est généralement pas différentiable, ce qui nous motive à considérer π_θ comme une classe de politiques stochastiques,qui permettent la différentiabilité

Policy Class

- Softmax polici

$$\pi_\theta(a \mid s) = \frac{\exp\left(\theta_{s,a}\right)}{\sum_{a'} \exp\left(\theta_{s,a'}\right)}, \quad \text{avec} \quad \Theta = \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$$

  - La fermeture de la classe Softmax contient toutes les politiques stationnaires et déterministes

- Log-linear policies

$$\pi_\theta(a \mid s) = \frac{\exp\left(\theta \cdot \phi_{s,a}\right)}{\sum_{a' \in \mathcal{A}} \exp\left(\theta \cdot \phi_{s,a'}\right)}$$

  - $\phi_{s,a} \in \mathbb{R}^d$: est un vecteur de caractéristiques de (s, a).

Policy Class

- Neural softmax policies

$$\pi_\theta(a \mid s) = \frac{\exp\left(f_\theta(s, a)\right)}{\sum_{a' \in \mathcal{A}} \exp\left(f_\theta\left(s, a'\right)\right)}$$

  - f_θ où θ représente les paramètres, peut être déterminée par un réseau de neurones.

Policy Gradient

- Considérons une distribution initiale μ, où s_0 est tiré selon cette distribution, et suivant une politique π, définissons la distribution de la trajectoire τ par :

$$\mathrm{Pr}_{\mu}^{\pi}(\tau) = \mu(s_0)\,\pi(a_0 \mid s_0)\,P(s_1 \mid s_0, a_0)\,\pi(a_1 \mid s_1)\cdots .$$

- la récompense totale actualisée d'une trajectoire:

$$R(\tau) := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

- $$V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \mathrm{Pr}_{\mu}^{\pi_\theta}}[R(\tau)].$$

Policy Gradient

Theorem

Il existe trois expressions différentes pour $\nabla_\theta V^{\pi_\theta}(\mu)$:

- *REINFORCE:*

$$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \mathrm{Pr}_\mu^{\pi_\theta}} \left[R(\tau) \sum_{t=0}^{\infty} \nabla \log \pi_\theta \left(a_t \mid s_t \right) \right]$$

- *Action value expression:*

$$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \mathrm{Pr}_\mu^{\pi_\theta}} \left[\sum_{t=0}^{\infty} \gamma^t Q^{\pi_\theta} \left(s_t, a_t \right) \nabla \log \pi_\theta \left(a_t \mid s_t \right) \right]$$

Policy Gradient

Theorem

- *Advantage expression:*

$$\nabla V^{\pi_\theta}(\mu) = \frac{1}{1-\gamma} \mathbb{E}_{s \sim d_\mu^{\pi_\theta}} \mathbb{E}_{a \sim \pi_\theta(\cdot|s)} \left[A^{\pi_\theta}(s,a) \nabla \log \pi_\theta(a \mid s)\right]$$

Avec

$$A^\pi(s,a) := Q^\pi(s,a) - V^\pi(s).$$

Appelé L'avantage d'une politiqueπ ($A^\pi(s,a) \leq 0$*then* $\pi = \pi^*$ *est*

$$d_\mu^\pi(s) = \mathbb{E}_{s_0 \sim \mu}\left[d_{s_0}^\pi(s)\right], d_{s_0}^\pi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \Pr^\pi(s_t = s \mid s_0).$$

$d_{s_0}^\pi(s)$: *la mesure de visite*

Policy Gradient

Preuve: indication

1) $$\nabla V^{\pi_\theta}(\mu) = \mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}}[R(\tau)] \quad = \nabla \sum_\tau R(\tau) \Pr_\mu^{\pi_\theta}(\tau)$$

2) $$\nabla V^{\pi_\theta}(s_0) = \nabla \sum_{a_0} \pi_\theta(a_0 \mid s_0) Q^{\pi_\theta}(s_0, a_0)$$

Gradient ascent

- **Remarque**
  $V^{\pi_\theta}(s)$ n'est pas concave lorsque l'on utilise le paramétrage Softmax.
- **Algorithme de Gradient ascent avec un pas fixe** η

$$\theta_{t+1} = \theta_t + \eta \nabla V^{\pi_{\theta_t}}(\mu).$$

Lemma

Supposons que $\theta \in \Theta$ et que V^{π_θ} est β-smooth et minoré par V^*. Supposons que nous utilisons un pas fixe η. Pour tout T, on a :

$$\min_{t \leq T} \|\nabla V^{\pi_{\theta_t}}(\mu)\|^2 \leq \frac{2\beta\left(V^*(\mu) - V^{\pi_{\theta_0}}(\mu)\right)}{T}.$$

Monte Carlo estimation and stochastic gradient ascent

- Un problème se pose : même si l'on connaît tous les paramètres du MDP, le calcul du gradient sera très coûteux en termes de calcul. Nous pouvons utiliser des estimations non biaisées de π en nous basant uniquement sur l'accès à notre modèle à l'aide de simulations. Autrement dit, en supposant que nous pouvons obtenir des trajectoires échantillonnées $\tau \sim \Pr_{\mu}^{\pi_\theta}$ (simuler des parcours possibles dans l'environnement défini par le modèle)
- Nous ignorons que τ est une séquence de longueur infinie

Pour une trajectoire τ (estimateur), nous définissons l'estimateur non biaisé du gradient:

$$\widehat{\nabla V^{\pi_\theta}}(\mu) := \sum_{t=0}^{\infty} \gamma^t \widehat{Q^{\pi_\theta}}(s_t, a_t) \nabla \log \pi_\theta (a_t \mid s_t)$$

Avec

$$\widehat{Q^{\pi_\theta}}(s_t, a_t) := \sum_{t'=t}^{\infty} \gamma^{t'-t} r(s_{t'}, a_{t'})$$

Lemma

(Unbiased gradient estimate) on a :

$$\mathbb{E}_{\tau \sim \Pr_\mu^{\pi_\theta}} \left[\widehat{\nabla V^{\pi_\theta}}(\mu) \right] = \nabla V^{\pi_\theta}(\mu)$$

stochastic gradient ascent algorithm

1. Initialiser θ_0.
2. Pour $t = 0, 1, \ldots$

(2.1) Échantillonner $\tau \sim \mathrm{Pr}_{\mu}^{\pi_\theta}$.

(2.2) Mettre à jour :

$$\theta_{t+1} = \theta_t + \eta_t \widehat{\nabla V^{\pi_\theta}}(\mu)$$

où η_t est la taille du pas et $\nabla \widehat{V^{\pi_\theta}(\mu)}$ est estimé avec τ.

stochastic gradient ascent algorithm

Lemma

(Stochastic Convergence to Stationary Points) Supposons que pour tout $\theta \in \Theta$, V^{π_θ} *soit* β*-smooth et borné inférieurement par* V^*. *Supposons que la variance soit bornée comme suit :*

$$\mathbb{E}\left[\left\|\widehat{\nabla V^{\pi_\theta}}(\mu) - \nabla V^{\pi_\theta}(\mu)\right\|^2\right] \leq \sigma^2$$

Pour $t \leq \beta\left(V^*(\mu) - V^{(0)}(\mu)\right)/\sigma^2$, *supposons que nous utilisions une taille de pas constante de* $\eta_t = 1/\beta$, *et par la suite, nous utilisons* $\eta_t = \sqrt{2/(\beta T)}$. *Pour tout* T, *nous avons :*

$$\min_{t \leq T} \mathbb{E}\left[\|\nabla V^{\pi_{\theta_t}}(\mu)\|^2\right] \leq \frac{2\beta\left(V^*(\mu) - V^{\pi_{\theta_0}}(\mu)\right)}{T} + \sqrt{\frac{2\sigma^2}{T}}.$$

stochastic gradient ascent algorithm

En pratique, σ est très élevée, c'est-à-dire que l'erreur non systématique est très grande dans l'estimateur que nous avons donné. Même s'il est sans biais, il n'est pas précis. Pour résoudre ce problème, nous utilisons une forme de réduction de la variance.

Soit $f : \mathcal{S} \to \mathbb{R}$.

1) Donner un estimateur de $V^{\pi_\theta}(\mu)$.

2)Échantillonner $\tau \sim \mathrm{Pr}_\mu^{\pi_\theta}$.

3) définie:

$$\widehat{Q^{\pi_\theta}}\left(s_t, a_t\right) := \sum_{t'=t}^{\infty} \gamma^{t'-t} r\left(s_{t'}, a_{t'}\right)$$

$$\widehat{\nabla V^{\pi_\theta}}(\mu) := \sum_{t=0}^{\infty} \gamma^t \left(\widehat{Q^{\pi_\theta}}\left(s_t, a_t\right) - f\left(s_t\right)\right) \nabla \log \pi_\theta\left(a_t \mid s_t\right)$$

Unbiased gradient estimate with Variance Reduction

Lemma

(Unbiased gradient estimate with Variance Reduction) Pour toute procédure utilisée pour construire la fonction de référence $f : \mathcal{S} \to \mathbb{R}$, si les échantillons utilisés pour construire f sont indépendants de la trajectoire τ, où $\widehat{Q^{\pi_\theta}}(s_t, a_t)$ est construit en utilisant τ, alors :

$$\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \left(\widehat{Q^{\pi_\theta}}(s_t, a_t) - f(s_t)\right) \nabla \log \pi_\theta (a_t \mid s_t)\right] = \nabla V^{\pi_\theta}(\mu)$$