Aprendizado por Reforço

AULA - 4

Gradientes de Política

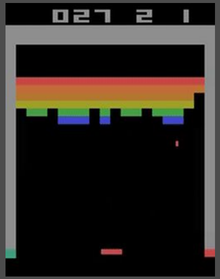
Retrospectiva do último episódio

- Diferença Temporal

$$V(s) \leftarrow V(s) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s)]$$

- Deep Q-Network

$$L(\theta) = \left(r + \gamma \max_{a'} Q(s', a'; \phi) - Q(s, a; \theta) \right)^2$$



Gradiente de Política

O que são Métodos de Gradiente de Política?

- Atualização da política de decisão DIRETAMENTE
  - Sem função de valor
- Gradiente usado para atualizar parâmetros θ de uma política π_θ

Como encontrar um gradiente “usável”?

- Saber qual o objetivo (o que maximizar ou minimizar)
- Escrever esse objetivo em termos dos parâmetros a serem otimizados

Objetivo de Reforço

- Maximizar recompensas ao longo do tempo
- Retorno / Retorno com Desconto

$$R_t = r_{t+1} + r_{t+2} + \cdots + r_T$$

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

Antes... Uma definição matemática

- Dado um caminho de estados e ações, eu posso definir matematicamente a probabilidade do mesmo

$$\tau = (s_1, a_1, s_2, \ldots s_T, a_T)$$

$$\rho_\theta(\tau) = \rho(s_1) \prod_{t=1}^{T} \pi_\theta(a_t | s_t) \rho(s_{t+1} | s_t, a_t)$$

Probabilidade de começar em s1

Objetivo

Retorno, com ou sem fator de desconto, do episódio inteiro

$$\theta^* = \max_{\theta} \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \underbrace{\left[\sum_t r(s_t, a_t) \right]}_{r(\tau)}$$

$$J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} [r(\tau)]$$

$$J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)}\left[r(\tau)\right] = \int \rho_\theta(\tau) r(\tau) d\tau$$

$$\nabla_\theta J(\theta) = \int \underline{\nabla_\theta \rho_\theta(\tau)} r(\tau) d\tau$$

$$\boxed{\rho_\theta(\tau) \nabla_\theta \log \rho_\theta(\tau) = \rho_\theta(\tau) \frac{\nabla_\theta \rho_\theta(\tau)}{\rho_\theta(\tau)} = \nabla_\theta \rho_\theta(\tau)}$$

$$\nabla_\theta J(\theta) = \int \underline{\rho_\theta(\tau) \nabla_\theta \log \rho_\theta(\tau)} r(\tau) d\tau$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\nabla_\theta \log \underline{\rho_\theta(\tau)} r(\tau)\right]$$

$$\nabla_\theta J(\theta) = \mathbb{E}\left[\nabla_\theta \log \rho_\theta(\tau) r(\tau)\right]$$

$$\rho_\theta(\tau) = \rho(s_1) \prod_{t=1}^{T} \pi_\theta(a_t|s_t)\rho(s_{t+1}|s_t, a_t)$$

$$\log \rho_\theta(\tau) = \log \rho(s_1) + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \log \rho(s_{t+1}|s_t, a_t)$$

$$\nabla_\theta \log \rho_\theta(\tau) = \nabla_\theta \left[\cancel{\log \rho(s_1)} + \sum_{t=1}^{T} \log \pi_\theta(a_t|s_t) + \cancel{\log \rho(s_{t+1}|s_t, a_t)}\right]$$

Independe de θ **Independe de θ**

$$\nabla_\theta J(\theta) = \mathbb{E}_{\tau \sim \rho_\theta(\tau)} \left[\left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t | s_t) \right) \left(\sum_{t=1}^{T} r(s_t, a_t) \right) \right]$$

$$\nabla_\theta J(\theta) \approx \underline{\frac{1}{N} \sum_{i=1}^{N}} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t} | s_{i,t}) \right) \underline{\left(\sum_{t=1}^{T} r(s_{i,t}, a_{i,t}) \right)}$$

A expectativa pode ser aproximada com a média de “muitas” amostras/caminhos

Retorno, com ou sem fator de desconto, do episódio inteiro

Gradiente de Política

- Caminhos com bons retornos tem suas ações incentivadas de acordo com suas probabilidades
  - Parâmetros são modificados para aumentar as probabilidades destas ações
- Caminhos com retornos ruins tem suas ações desencorajadas
  - Parâmetros são modificados para aumentar as probabilidades destas ações

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_t | s_t) \right) \left(\sum_{t=1}^{T} r(s_t, a_t) \right)$$

Algoritmo REINFORCE

- Amostras de episódios completos
- Sem *Bootstrapping* ou Estimativa de Valor

REINFORCE algorithm:

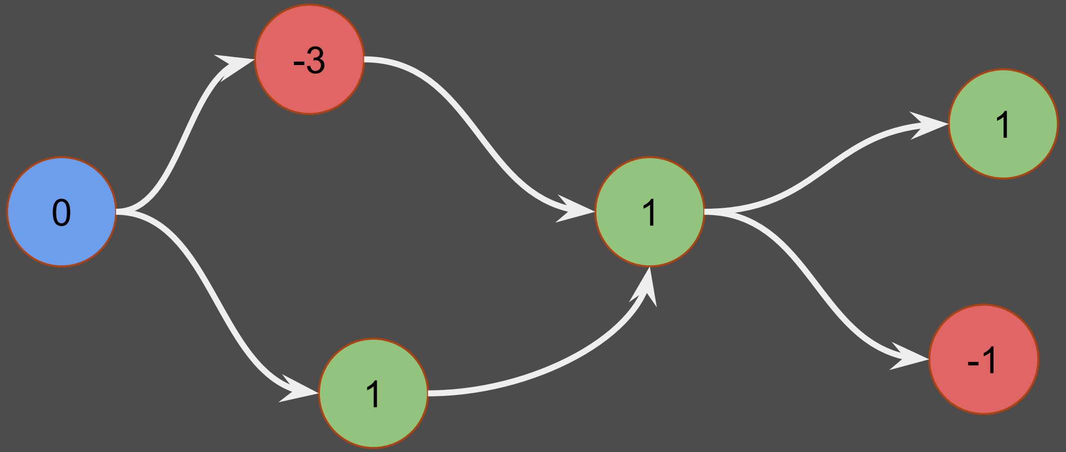
1. sample $\{\tau^i\}$ from $\pi_\theta(\mathbf{a}_t|\mathbf{s}_t)$ (run the policy)
2. $\nabla_\theta J(\theta) \approx \sum_i \left(\sum_t \nabla_\theta \log \pi_\theta(\mathbf{a}_t^i|\mathbf{s}_t^i)\right)\left(\sum_t r(\mathbf{s}_t^i, \mathbf{a}_t^i)\right)$
3. $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$

- Não foi originalmente pensado para grandes espaços de estados

Problemas do Gradiente

Variância

- Ações são encorajadas e desencorajadas a partir do retorno do episódio inteiro
- Ações não necessariamente ligadas a uma recompensa ruim, podem estar sendo desencorajadas
  - Existe correlação de causalidade para o futuro, mas não para o passado



Variância

- Uma ação pode ter muitas atualizações diferentes por conta de ações passadas do agente
- Solução:
- Considerar apenas o retorno com recompensas futuras (a partir de t)

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\left(\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})\right)\left(\sum_{t'=t}^{T} r(s_{i,t'}, a_{i,t'})\right)$$

$$\boxed{\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) R_{i,t}}$$

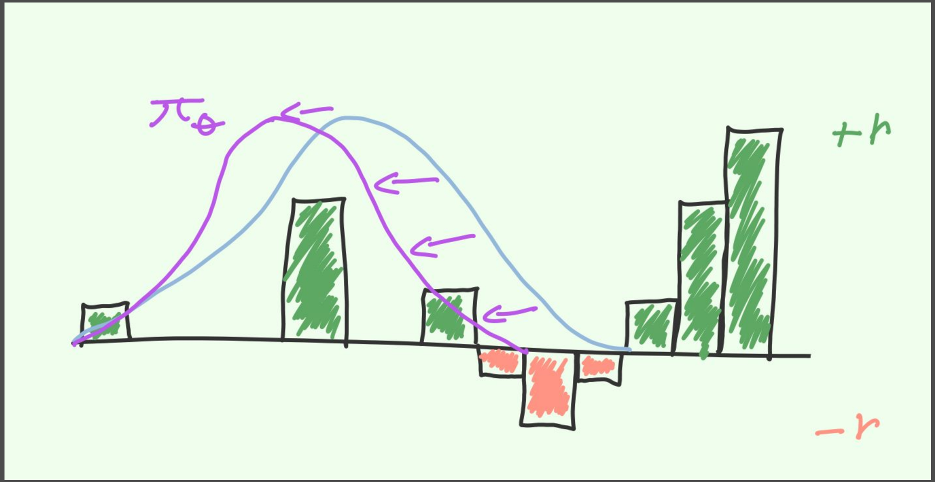
Outros problemas...

- E se os melhores retornos forem zero?
    - Probabilidades não são atualizadas
- E se continuar reforçando uma ação boa?
    - Supervalorização de ações ainda atrapalham outras probabilidades porque mexem nos mesmos parâmetros

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t}) R_{i,t}$$

Outros problemas...

- O gradiente leva a distribuição para longe de retornos ruins. Porém, não sabemos se, no espaço de soluções, há retornos melhores após estes ruins.
- Famoso problema dos *mínimos locais* (neste caso, máximo)



Calma...
Não desista ainda

Baselines

Resolvendo seus problemas de Gradiente de Política desde 1987

Linha de Base para o Retorno

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})[R_{i,t} - b]$$

$$b = \frac{1}{N}\sum_{i=1}^{N} r(\tau)$$

- Se *b* for uma média dos retornos, as probabilidades serão atualizadas de acordo com a diferença de um retorno R e a média dos caminhos coletados.

Resolvendo Problemas

- Um retorno de zero que for acima da média, terá suas ações incentivadas
- Retornos comuns não terão suas ações “supervalorizadas”
- Se uma recompensa negativa é “esperada” a distribuição não se distanciará tanto, diminuindo a chance de cair em máximos locais


ao vivo
PODE ISSO ARNALDO?

$$\nabla_\theta J(\theta) \approx \frac{1}{N}\sum_{i=1}^{N}\sum_{t=1}^{T}\nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})[R_{i,t} - b]$$

$$\rho_\theta(\tau)\nabla_\theta \log \rho_\theta(\tau) = \rho_\theta(\tau)\frac{\nabla_\theta \rho_\theta(\tau)}{\rho_\theta(\tau)} = \nabla_\theta \rho_\theta(\tau)$$

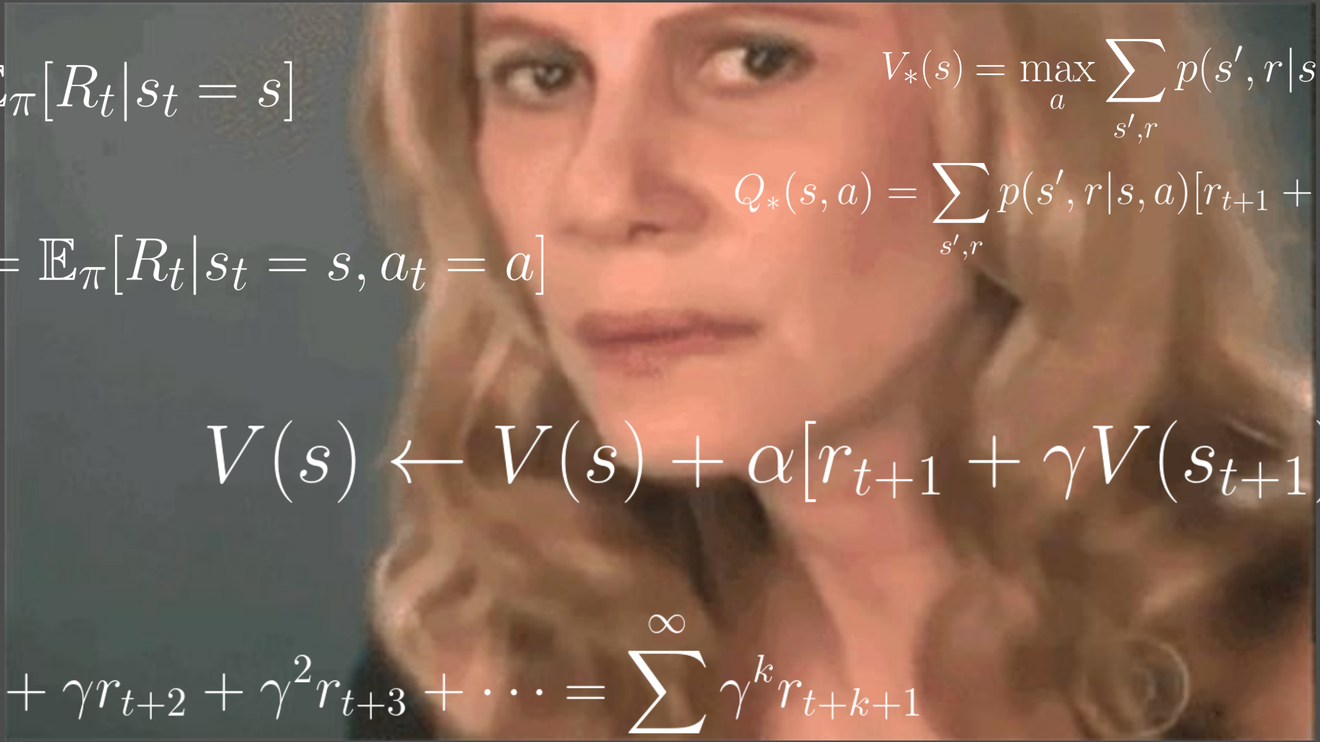
$$E[\nabla_\theta \log p_\theta(\tau)b] = \int p_\theta(\tau)\nabla_\theta \log p_\theta(\tau)b\, d\tau = \int \nabla_\theta p_\theta(\tau)b\, d\tau = b\nabla_\theta \int p_\theta(\tau)d\tau = b\nabla_\theta 1 = 0$$

Se b não depender das ações, ele não afeta a expectativa

Quais seriam bons candidatos para b?

- Valor fixo?
- Média dos Retornos coletados?
- Média dos retornos do estado st especificamente?
- Uma estimativa do retorno do estado st?

Uma estimativa do retorno do estado st?



$$V_\pi(s) = \mathbb{E}_\pi[R_t | s_t = s]$$

$$V_*(s) = \max_a \sum_{s',r} p(s', r | s, a)[r_{t+1} + \gamma V_*(s')]$$

$$Q_*(s, a) = \sum_{s',r} p(s', r | s, a)[r_{t+1} + \gamma \max_{a'} Q_*(s_{t+1}, a')]$$

$$Q_\pi(s, a) = \mathbb{E}_\pi[R_t | s_t = s, a_t = a]$$

$$V(s) \leftarrow V(s) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s)]$$

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \cdots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}$$

Esse estimador pode ter seus próprios parâmetros

$$\nabla_\theta J(\theta) \approx \frac{1}{N} \sum_{i=1}^{N} \sum_{t=1}^{T} \nabla_\theta \log \pi_\theta(a_{i,t}|s_{i,t})[R_{i,t} - V_\phi(s_t)]$$

Ah... Isso é Actor Critic?

NÃO

Qual a diferença?

- Algoritmos de *Actor-Critic* fazem *Bootstrapping* do Retorno

$$R_t = r_{t+1} + \gamma V_\phi(s)$$

- Chama-se **Crítico** porque o modelo estima o quão bem a política irá performar no futuro, dado também que ele aprende com as experiências coletadas por tal política

Agora podemos enfrentar os métodos de *Actor-Critic*