INSTITUTO DE MATEMATICA E ESTATÍSTICA
UNIVERSIDADE DE SÃO PAULO

VERÃO
IME - USP

XLIX Programa de Verão (2020) - Introdução ao Aprendizado por Reforço

Redução de Variância e Funções Valor

Thiago Pereira Bueno
tbueno@ime.usp.br

IME - USP, 13/02/2019
LIAMF: Grupo PAR (Planejamento e Aprendizado por Reforço)

Aula 3

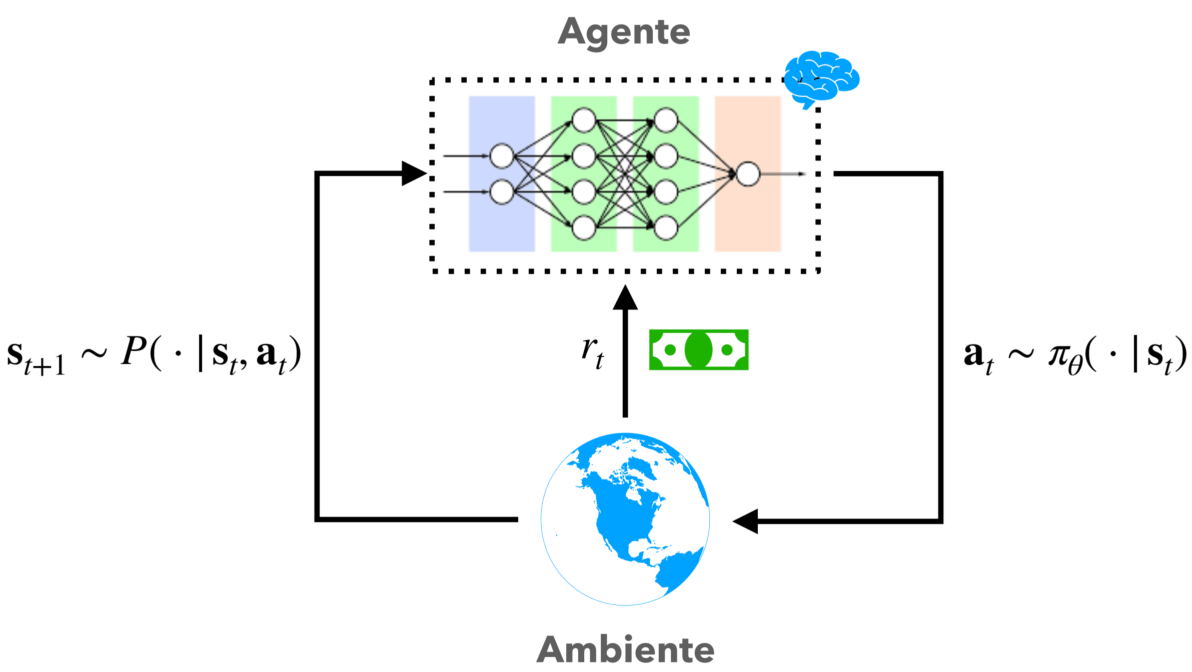
Agenda

1. Policy Gradients e REINFORCE: revisão
2. Propriedades do Score Function
3. Redução de Variância via *Reward-to-Go*
4. Funções Valor e *Baseline:* escala de referência para retornos

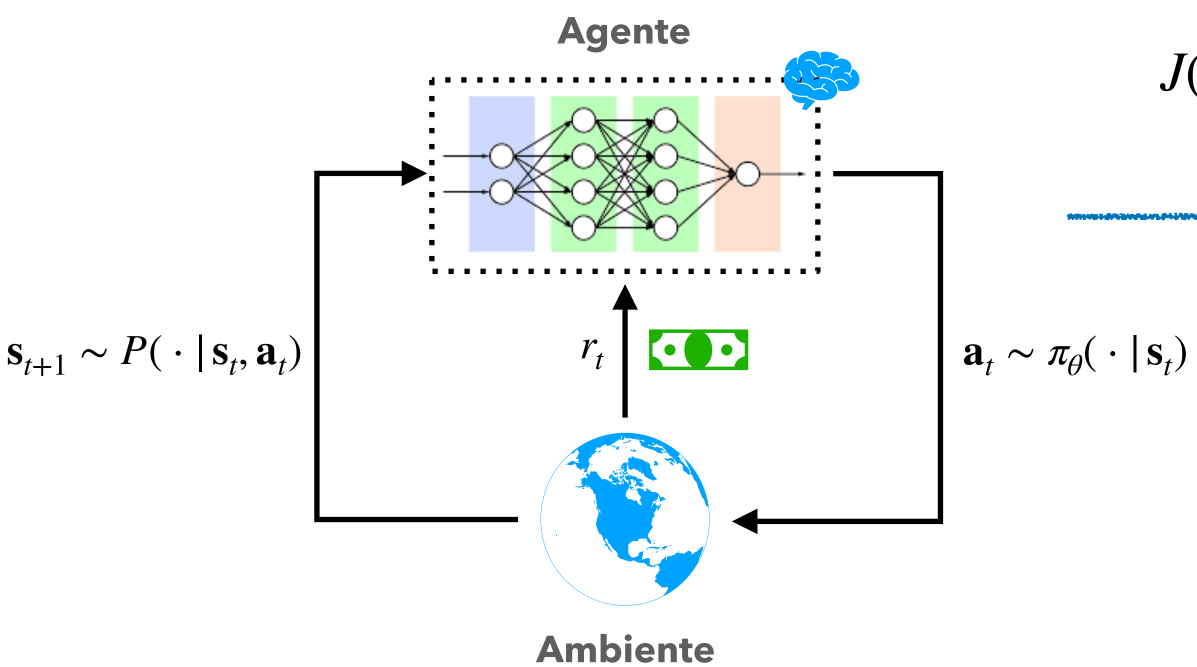
Objetivos

- Relacionar as propriedades do estimador REINFORCE com a performance do agente
- Entender e implementar técnicas estatísticas para redução de ruído / variância
- Aprender um aproximador paramétrico para estimar retornos esperados

Policy Gradients e REINFORCE: revisão (1/3)

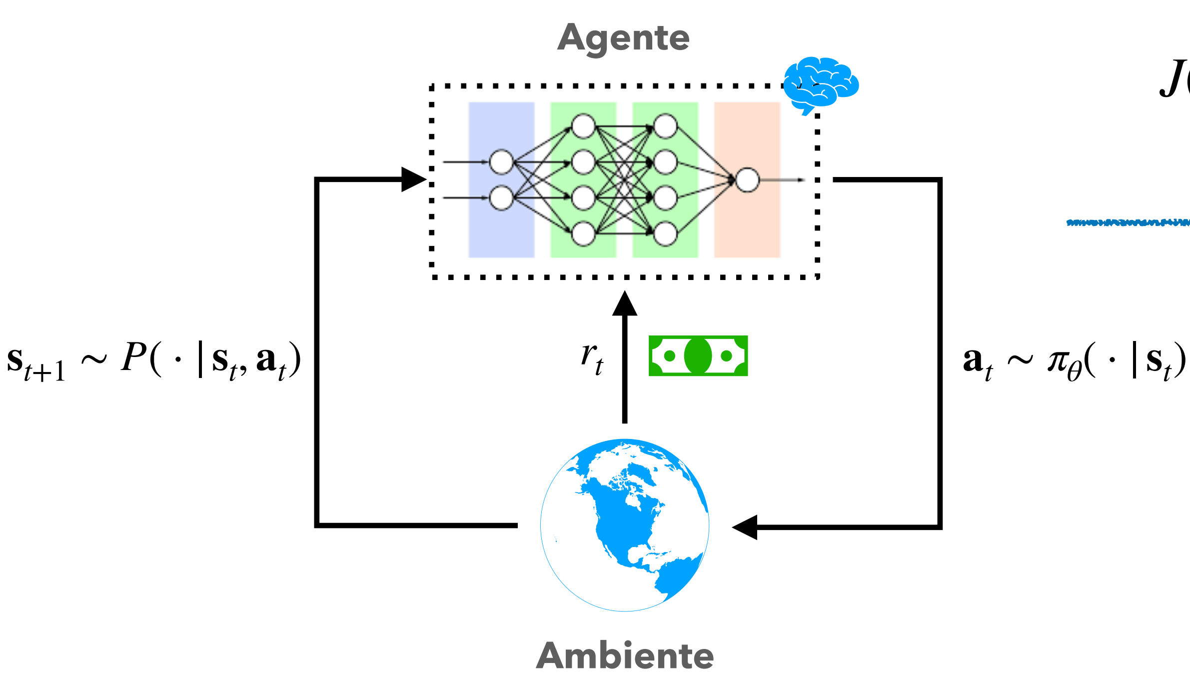


Policy Gradients e REINFORCE: revisão (1/3)

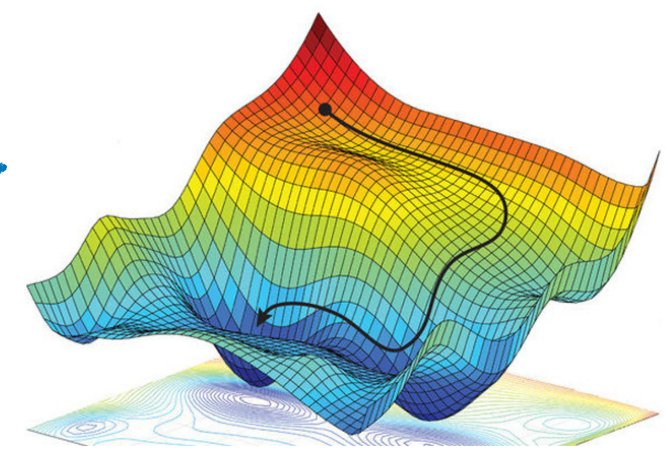


$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t\right]$$

Policy Gradients e REINFORCE: revisão (1/3)

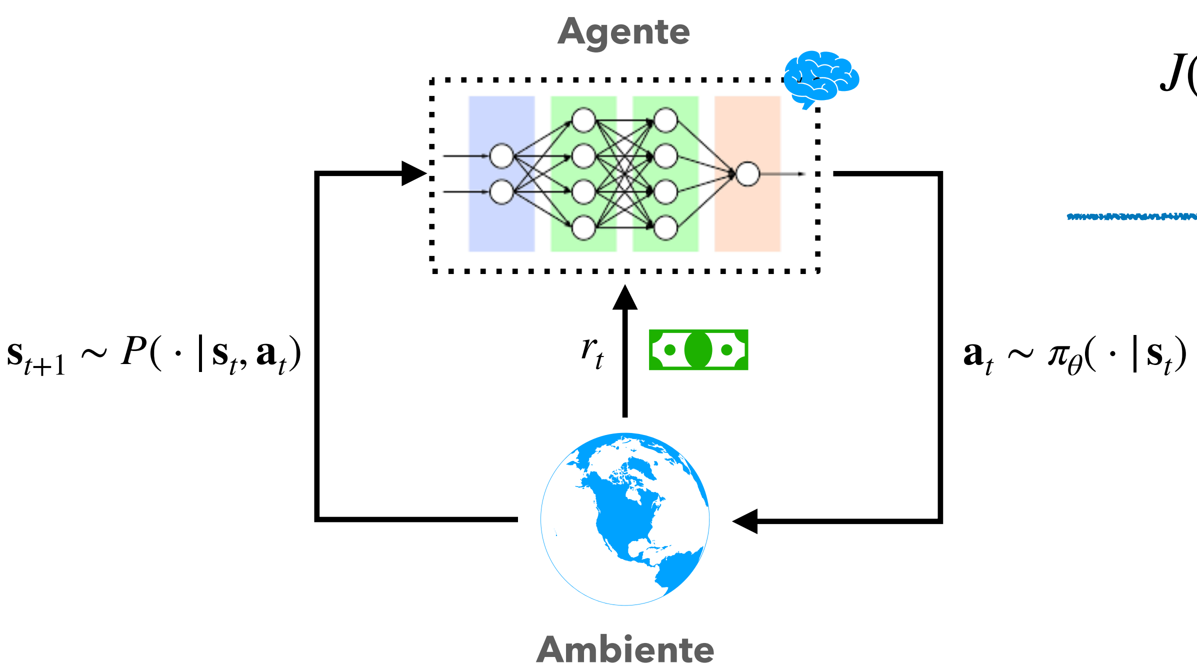


$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t\right]$$

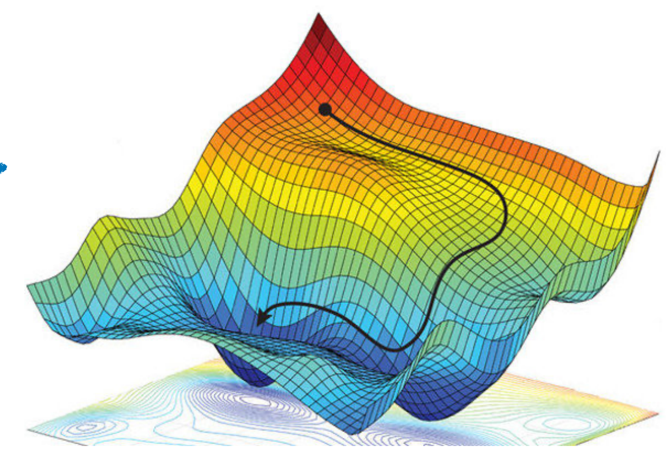


$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

Policy Gradients e REINFORCE: revisão (1/3)



$$J(\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t\right]$$



$$\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$$

Policy Gradient

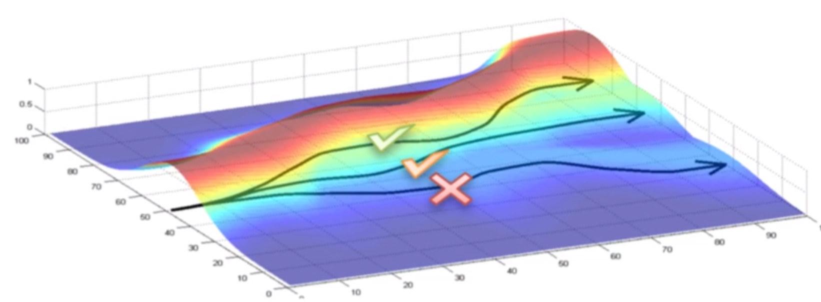
$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t \mid s_t) R(\tau)\right]$$

Policy Gradients e REINFORCE: revisão (2/3)

$$\nabla_\theta J(\theta) = \mathop{\mathbb{E}}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \underbrace{\nabla_\theta \log \pi_\theta(a_t \mid s_t)}_{\textit{Score Function}}\, \underbrace{R(\tau)}_{\textbf{Retorno}}\right]$$

REINFORCE = Tentativa & Erro

$R(\tau) > 0$	$\uparrow \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)$	**Reforço positivo**
$R(\tau) < 0$	$\downarrow \pi_\theta(\mathbf{a}_t \mid \mathbf{s}_t)$	**Reforço negativo**



Policy Gradients e REINFORCE: revisão (3/3)

Algoritmo 1 REINFORCE

Entrada: parâmetros da política, θ

1: **enquanto** não satisfeito **faça**
2: Colete trajetórias com a política atual, $\tau_1, \tau_2, \ldots, \tau_N \sim \pi_\theta$
3: Calcule os retornos de cada trajetória, $R_k = \sum_{t=1}^{T} r_t^k$
4: Estime o *Policy Gradient*

$$\nabla_\theta J(\theta) \approx \boxed{\frac{1}{N} \sum_{k=1}^{N} \sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(a_t^k \mid s_t^k) R_k} \longleftarrow$$

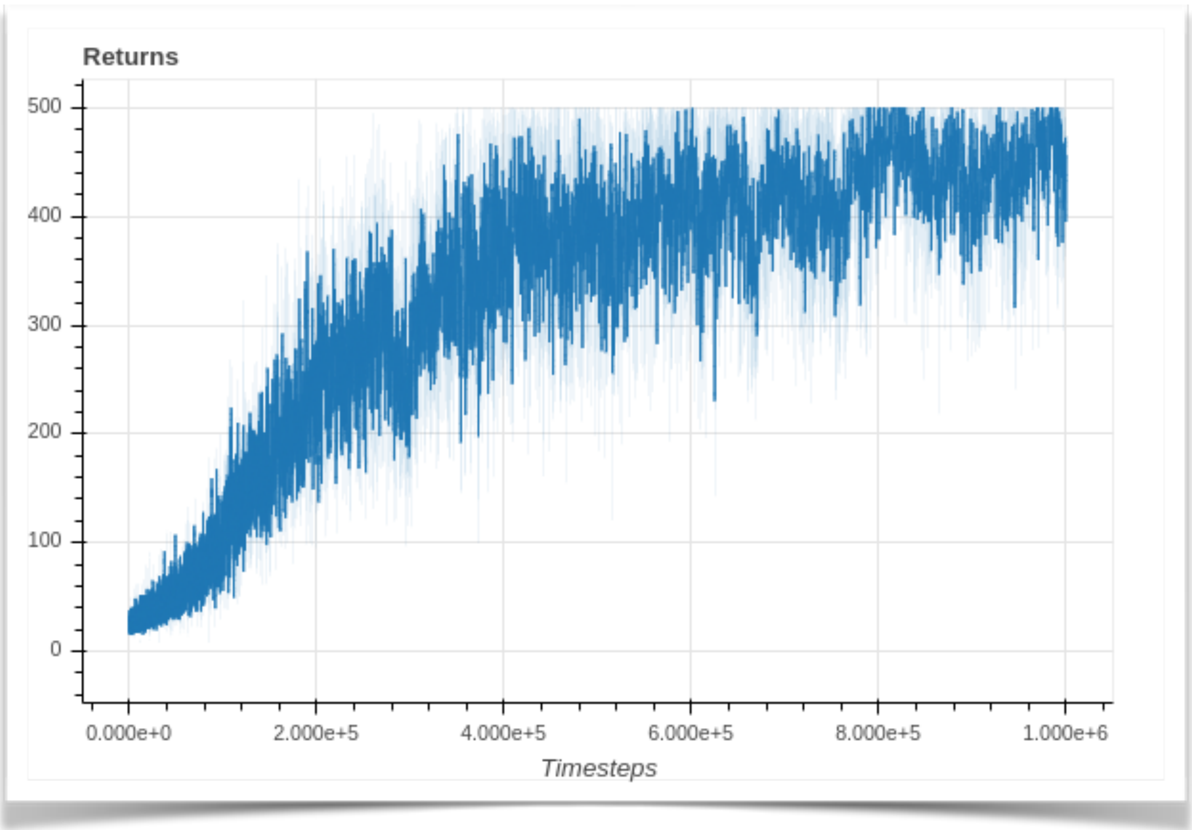
Monte-Carlo Sampling

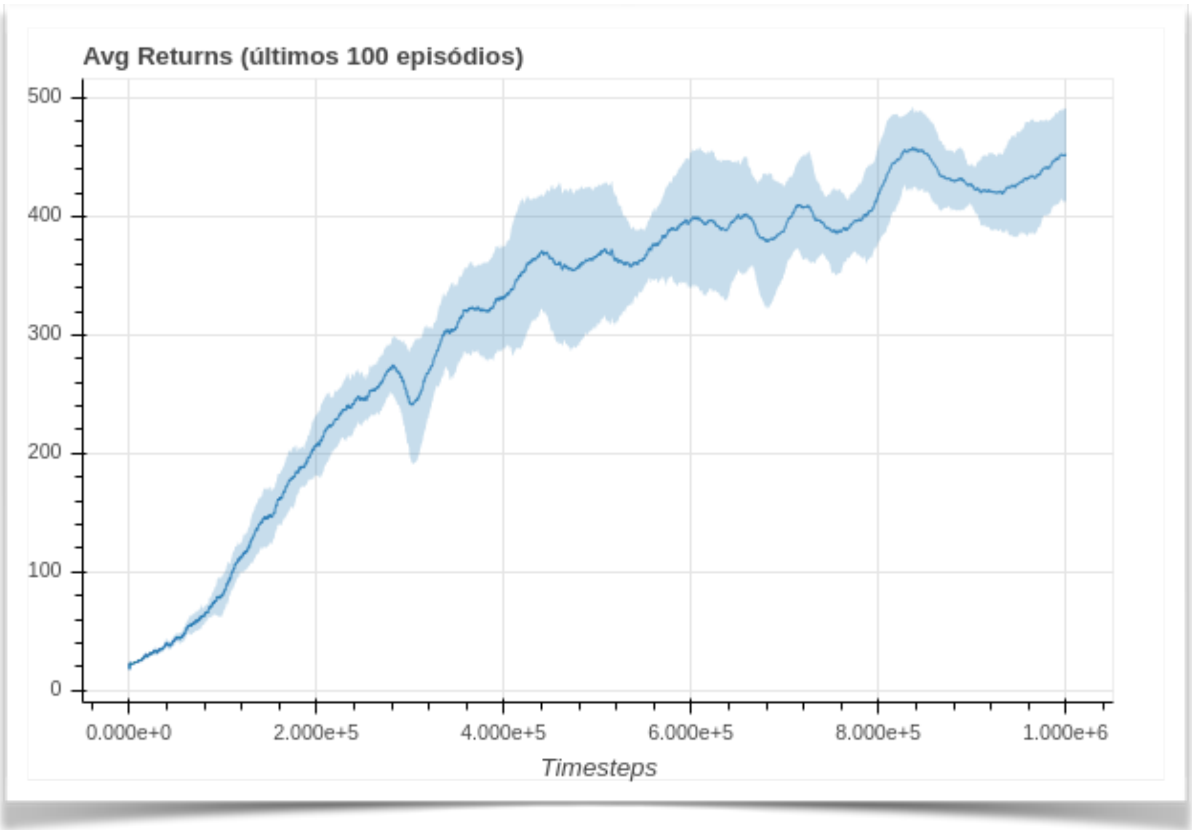
5: Atualize os parâmetros da política, $\theta \leftarrow \theta + \alpha \nabla_\theta J(\theta)$
6: **fim enquanto**
7: **devolve** π_θ

Na aula de hoje vamos nos concentrar em ***estudar as propriedades*** *do estimador* **REINFORCE**

Como acelerar o treinamento de políticas?

CartPole-v1 (média e desvio padrão para 10 *trials)*





- Quais as propriedades do ***Policy Gradient*** que impactam no treinamento?
  - Qual a relação da **variância do estimador** com o ruído nos experimentos?
  - Como uma estimativa enviesada do gradiente pode interferir na performance final ?

Propriedades do *Score Function*

$$\mathbb{E}_{x\sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\textit{Score Function}}\right] = \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Propriedades do *Score Function*

$$\mathbb{E}_{x\sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\textit{Score Function}}\right] = \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx$$

$$= \int p_\theta(x)\frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}dx$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Propriedades do *Score Function*

$$
\begin{aligned}
\mathbb{E}_{x\sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\text{Score Function}}\right] &= \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx \\
&= \int p_\theta(x)\frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}dx \\
&= \int \nabla_\theta p_\theta(x)dx
\end{aligned}
$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Propriedades do *Score Function*

$$
\begin{aligned}
\mathbb{E}_{x \sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\text{Score Function}}\right] &= \int p_\theta(x) \nabla_\theta \log p_\theta(x) dx \\
&= \int p_\theta(x) \frac{\nabla_\theta p_\theta(x)}{p_\theta(x)} dx \\
&= \int \nabla_\theta p_\theta(x) dx \\
&= \nabla_\theta \int p_\theta(x) dx
\end{aligned}
$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Propriedades do *Score Function*

$$
\begin{aligned}
\mathbb{E}_{x\sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\text{Score Function}}\right] &= \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx \\
&= \int p_\theta(x)\frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}dx \\
&= \int \nabla_\theta p_\theta(x)dx \\
&= \nabla_\theta \int p_\theta(x)dx \\
&= \nabla_\theta 1
\end{aligned}
$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Propriedades do *Score Function*

$$
\begin{aligned}
\mathbb{E}_{x\sim p_\theta}\left[\underbrace{\nabla_\theta \log p_\theta(x)}_{\textit{Score Function}}\right] &= \int p_\theta(x)\nabla_\theta \log p_\theta(x)dx \\
&= \int p_\theta(x)\frac{\nabla_\theta p_\theta(x)}{p_\theta(x)}dx \\
&= \int \nabla_\theta p_\theta(x)dx \\
&= \nabla_\theta \int p_\theta(x)dx \\
&= \nabla_\theta 1 \\
&= 0
\end{aligned}
$$

Teorema: o valor esperado do *score function* de uma distribuição paramétrica é zero.

Redução de Variância via *Reward-to-Go* (1/3)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right]$$

Redução de Variância via *Reward-to-Go* (1/3)

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\ &= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right]\end{aligned}$$

Redução de Variância via *Reward-to-Go* (1/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=0}^{T-1} r_{t'}\right)\right]
\end{aligned}
$$

Redução de Variância via *Reward-to-Go* (1/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=0}^{T-1} r_{t'}\right)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=0}^{t-1} r_{t'} + \sum_{t'=t}^{T-1} r_{t'}\right)\right]
\end{aligned}
$$

Redução de Variância via *Reward-to-Go* (1/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\sum_{t'=0}^{T-1} r_{t'}\right)\right] \\
&= \sum_{t=0}^{T-1} \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \left(\underbrace{\sum_{t'=0}^{t-1} r_{t'}}_{\textbf{passado}} + \underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textbf{futuro}}\right)\right]
\end{aligned}
$$

Redução de Variância via *Reward-to-Go* (2/3)

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\underbrace{\sum_{t'=0}^{t-1} r_{t'}}_{\text{passado}}\right] =$$

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (2/3)

$$\mathbb{E}_{\tau \sim \pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=0}^{t-1} r_{t'}}_{\text{passado}}\right] = \mathbb{E}_{\tau_{0:t} \sim \pi_\theta}\left[\mathbb{E}_{\mathbf{a}_t \sim \pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \sum_{t'=0}^{t-1} r_{t'} \,\middle|\, \tau_{0:t}\right]\right]$$

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (2/3)

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=0}^{t-1} r_{t'}}_{\text{passado}}\right] = \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \sum_{t'=0}^{t-1} r_{t'} \;\middle|\; \tau_{0:t}\right]\right]$$

$$= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right) \mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)|\tau_{0:t}\right]\right]$$

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (2/3)

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=0}^{t-1} r_{t'}\right] = \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=0}^{t-1} r_{t'} \,\middle|\, \tau_{0:t}\right]\right]$$

$$= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right)\boxed{\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)|\tau_{0:t}\right]}\right]$$

passado

Esperança do Score Function

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (2/3)

$$
\begin{aligned}
\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=0}^{t-1} r_{t'}\right] &= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=0}^{t-1} r_{t'} \;\middle|\; \tau_{0:t}\right]\right] \\
&= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right)\boxed{\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)|\tau_{0:t}\right]}\right] \\
&= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right)\cdot 0\right]
\end{aligned}
$$

passado

Esperança do Score Function

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (2/3)

$$\mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\underbrace{\sum_{t'=0}^{t-1} r_{t'}}_{\text{passado}}\right] = \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=0}^{t-1} r_{t'} \,\middle|\, \tau_{0:t}\right]\right]$$

$$= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right)\underbrace{\boxed{\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)|\tau_{0:t}\right]}}_{\text{Esperança do Score Function}}\right]$$

$$= \mathbb{E}_{\tau_{0:t}\sim\pi_\theta}\left[\left(\sum_{t'=0}^{t-1} r_{t'}\right)\cdot 0\right]$$

$$= 0$$

- *Lembre-se que o **valor esperado do Score Function** de uma distribuição é **zero**.*

Redução de Variância via *Reward-to-Go* (3/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} \left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textit{Reward-to-Go}}\right)\right]
\end{aligned}
$$

Redução de Variância via *Reward-to-Go* (3/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textit{Reward-to-Go}}\right)\right]
\end{aligned}
$$

1. O principal desafio do *Policy Gradient* é o número de trajetórias para se obter uma boa estimativa

Redução de Variância via *Reward-to-Go* (3/3)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textit{Reward-to-Go}}\right)\right]$$

1. O principal desafio do *Policy Gradient* é o número de trajetórias para se obter uma boa estimativa
2. A fórmula inicial continha termos que "reforçavam" ações de acordo com recompensas passadas

Redução de Variância via *Reward-to-Go* (3/3)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textit{Reward-to-Go}}\right)\right]$$

1. O principal desafio do *Policy Gradient* é o número de trajetórias para se obter uma boa estimativa
2. A fórmula inicial continha termos que “reforçavam” ações de acordo com recompensas passadas
3. Todos esses termos tem média zero, embora contribuam para a variância do estimador (i.e., ruído)

Redução de Variância via *Reward-to-Go* (3/3)

$$
\begin{aligned}
\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) R(\tau)\right] \\
&= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} \left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) \underbrace{\sum_{t'=t}^{T-1} r_{t'}}_{\textit{Reward-to-Go}}\right)\right]
\end{aligned}
$$

1. O principal desafio do *Policy Gradient* é o número de trajetórias para se obter uma boa estimativa
2. A fórmula inicial continha termos que "reforçavam" ações de acordo com recompensas passadas
3. Todos esses termos tem média zero, embora contribuam para a variância do estimador (i.e., ruído)
4. Ao remover esses termos do estimador, podemos reduzir o número de episódios necessários!

Funções *Baseline*: referência para retornos (1/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underline{b(\mathbf{s}_t)}\right)\right)\right]$$

baseline

Funções *Baseline*: referência para retornos (1/2)

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right] \\ &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underbrace{b(\mathbf{s}_t)}_{\textit{baseline}}\right)\right)\right]\end{aligned}$$

Intuição:

Funções *Baseline*: referência para retornos (1/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$

$$= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underbrace{b(\mathbf{s}_t)}_{\textit{baseline}}\right)\right)\right]$$

Intuição:

- A função *baseline* serve como um valor referência para o reforço dado pelo *reward-to-go*

Funções *Baseline*: referência para retornos (1/2)

$$\begin{aligned}\nabla_\theta J(\pi_\theta) &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right] \\ &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underbrace{b(\mathbf{s}_t)}_{\textit{baseline}}\right)\right)\right]\end{aligned}$$

Intuição:

- A função *baseline* serve como um valor referência para o reforço dado pelo *reward-to-go*

Note que sem essa referência ...

Funções *Baseline*: referência para retornos (1/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$
$$= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underbrace{b(\mathbf{s}_t)}_{\textit{baseline}}\right)\right)\right]$$

Intuição:

- A função *baseline* serve como um valor referência para o reforço dado pelo *reward-to-go*

Note que sem essa referência ...

- A probabilidade de uma boa ação pode ser diminuída se o retorno do episódio for negativo

Funções *Baseline*: referência para retornos (1/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underline{b(\mathbf{s}_t)}\right)\right)\right]$$

baseline

Intuição:

- A função *baseline* serve como um valor referência para o reforço dado pelo *reward-to-go*

Note que sem essa referência ...

- A probabilidade de uma boa ação pode ser diminuída se o retorno do episódio for negativo
- Ao longo do treinamento, piores ações serão mais desencorajadas; isso irá encorajar **indiretamente** as boas ações;

Funções *Baseline*: referência para retornos (1/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$

$$= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underbrace{b(\mathbf{s}_t)}_{\textit{baseline}}\right)\right)\right]$$

Intuição:

- A função *baseline* serve como um valor referência para o reforço dado pelo *reward-to-go*

Note que sem essa referência ...

- A probabilidade de uma boa ação pode ser diminuída se o retorno do episódio for negativo
- Ao longo do treinamento, piores ações serão mais desencorajadas; isso irá encorajar **indiretamente** as boas ações;
- No entanto, isso irá **retardar** consideravelmente o treinamento!

Funções *Baseline*: referência para retornos (2/2)

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\sum_{t'=t}^{T-1} r_{t'}\right)\right]$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1}\left(\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t)\left(\left(\sum_{t'=t}^{T-1} r_{t'}\right) - \underline{b(\mathbf{s}_t)}\right)\right)\right]$$

baseline

Prova (sketch): *use o teorema do valor esperado do score function para provar o lema:*

$$\mathbb{E}_{\mathbf{a}_t\sim\pi_\theta(\cdot|\mathbf{s}_t)}\left[\nabla_\theta \log \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) b(\mathbf{s}_t)\right] = 0$$

Funções Valor como Baseline (1/2)

Uma boa referência para o *retorno de um episódio* é dada pela **Função Valor**

$$
\begin{aligned}
b(\mathbf{s}) &= V^{\pi_\theta}(\mathbf{s}) \\
&= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t \;\middle|\; \mathbf{s}_0 = \mathbf{s} \right]
\end{aligned}
$$

Funções Valor como Baseline (1/2)

Uma boa referência para o *retorno de um episódio* é dada pela **Função Valor**

$$
\begin{aligned}
b(\mathbf{s}) &= V^{\pi_\theta}(\mathbf{s}) \\
&= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t \;\middle|\; \mathbf{s}_0 = \mathbf{s} \right]
\end{aligned}
$$

Note que naturalmente uma trajetória $\tau = (\mathbf{s}_0, \mathbf{a}_0, r_1, \cdots, \mathbf{s}_{T-1}, \mathbf{a}_{T-1}, r_T)$ **:**

Funções Valor como Baseline (1/2)

Uma boa referência para o *retorno de um episódio* é dada pela **Função Valor**

$$
\begin{aligned}
b(\mathbf{s}) &= V^{\pi_\theta}(\mathbf{s}) \\
&= \mathbb{E}_{\tau \sim \pi_\theta}\left[\sum_{t=0}^{T-1} r_t \;\middle|\; \mathbf{s}_0 = \mathbf{s} \right]
\end{aligned}
$$

Note que naturalmente uma trajetória $\tau = (\mathbf{s}_0, \mathbf{a}_0, r_1, \cdots, \mathbf{s}_{T-1}, \mathbf{a}_{T-1}, r_T)$ **:**

- pode ser considerada "**boa**" se $R(\tau' \,|\, \mathbf{s}_0 = \mathbf{s}) > \mathbb{E}_{\tau \sim \pi_\theta}[R(\tau) \,|\, \mathbf{s}_0 = \mathbf{s}]$

Funções Valor como Baseline (1/2)

Uma boa referência para o *retorno de um episódio* é dada pela **Função Valor**

$$\begin{aligned} b(\mathbf{s}) &= V^{\pi_\theta}(\mathbf{s}) \\ &= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=0}^{T-1} r_t \;\middle|\; \mathbf{s}_0 = \mathbf{s}\right] \end{aligned}$$

Note que naturalmente uma trajetória $\tau = (\mathbf{s}_0, \mathbf{a}_0, r_1, \cdots, \mathbf{s}_{T-1}, \mathbf{a}_{T-1}, r_T)$:

- pode ser considerada "**boa**" se $R(\tau' \mid \mathbf{s}_0 = \mathbf{s}) > \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau) \mid \mathbf{s}_0 = \mathbf{s}]$
- pode ser considerada "**ruim**" se $R(\tau' \mid \mathbf{s}_0 = \mathbf{s}) < \mathbb{E}_{\tau\sim\pi_\theta}[R(\tau) \mid \mathbf{s}_0 = \mathbf{s}]$

Funções Valor como Baseline (2/2)

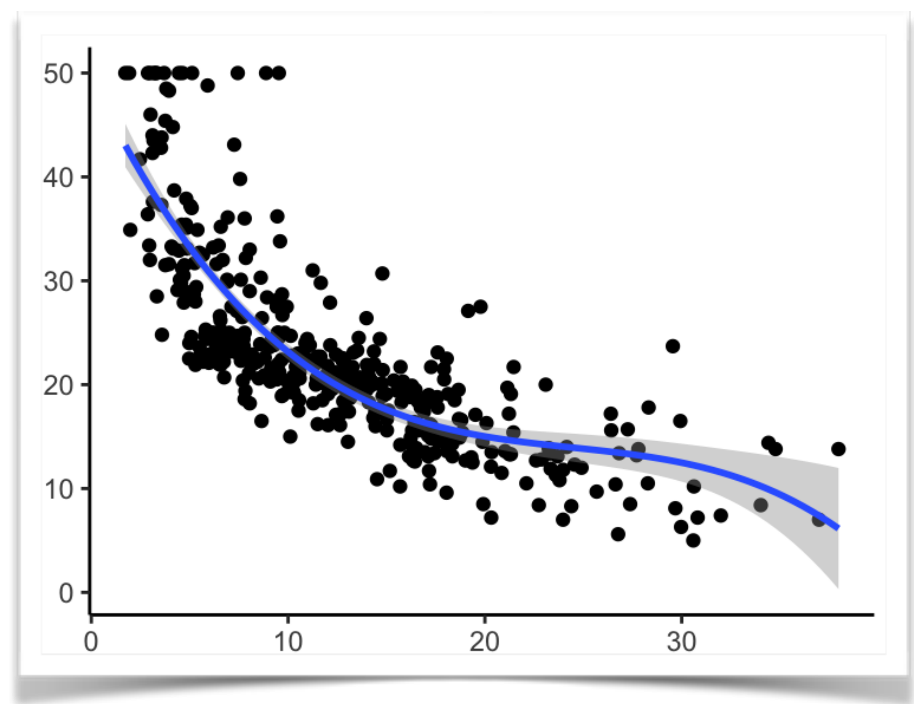
Problema: não temos acesso direto à função Valor de um estado!

Solução:

1. **Simular trajetórias** e estimar retornos de episódios via **Monte-Carlo**
2. Aproximar a função Valor via **regressão** (e.g., à la *Supervised Learning)*

$$\phi_k = \arg\min_{\phi} \mathbb{E}\left[\left(V_\phi(\mathbf{s}_t) - \sum_{t'=t}^{T-1} r'_t\right)^2\right]$$

Mean Squared Error



Referências

(1) **OpenAI Spinning Up**

- https://spinningup.openai.com/en/latest/spinningup/extra_pg_proof1.html
- https://spinningup.openai.com/en/latest/spinningup/rl_intro3.html#baselines-in-policy-gradients

(2) **Deep RL Bootcamp Lecture 4A: Policy Gradients**

- https://www.youtube.com/watch?v=S_gwYj1Q-44

(3) **Reinforcement Learning: An Introduction** (Sutton & Barto 2018, 2nd Edition)

- Seções 13.3 e 13.4