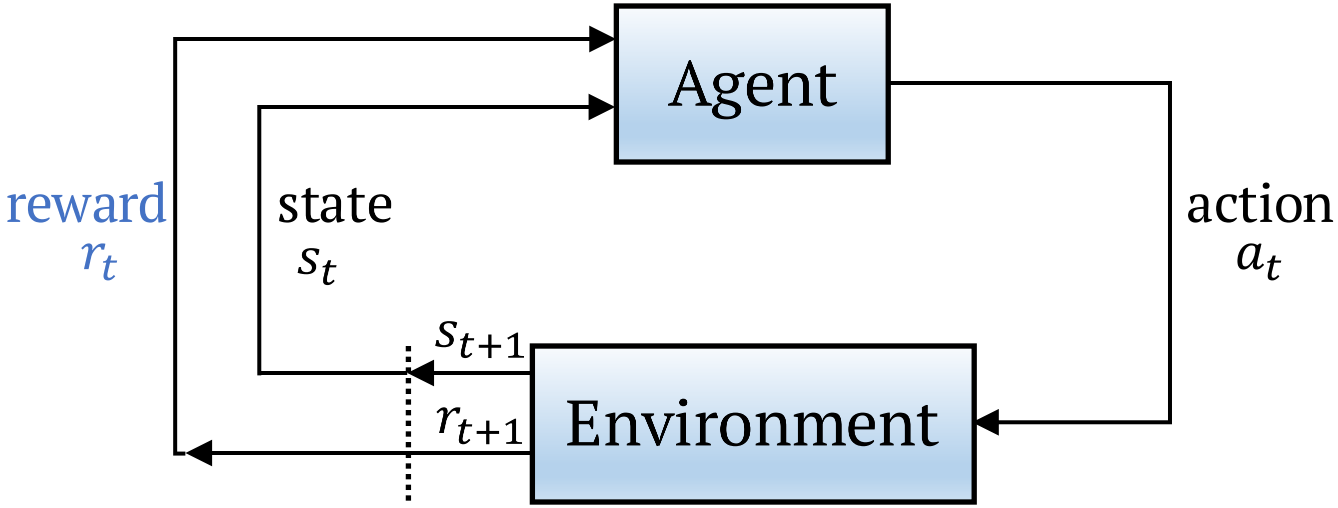
Natural Policy Gradient Primal-Dual Method for Constrained Markov Decision Processes

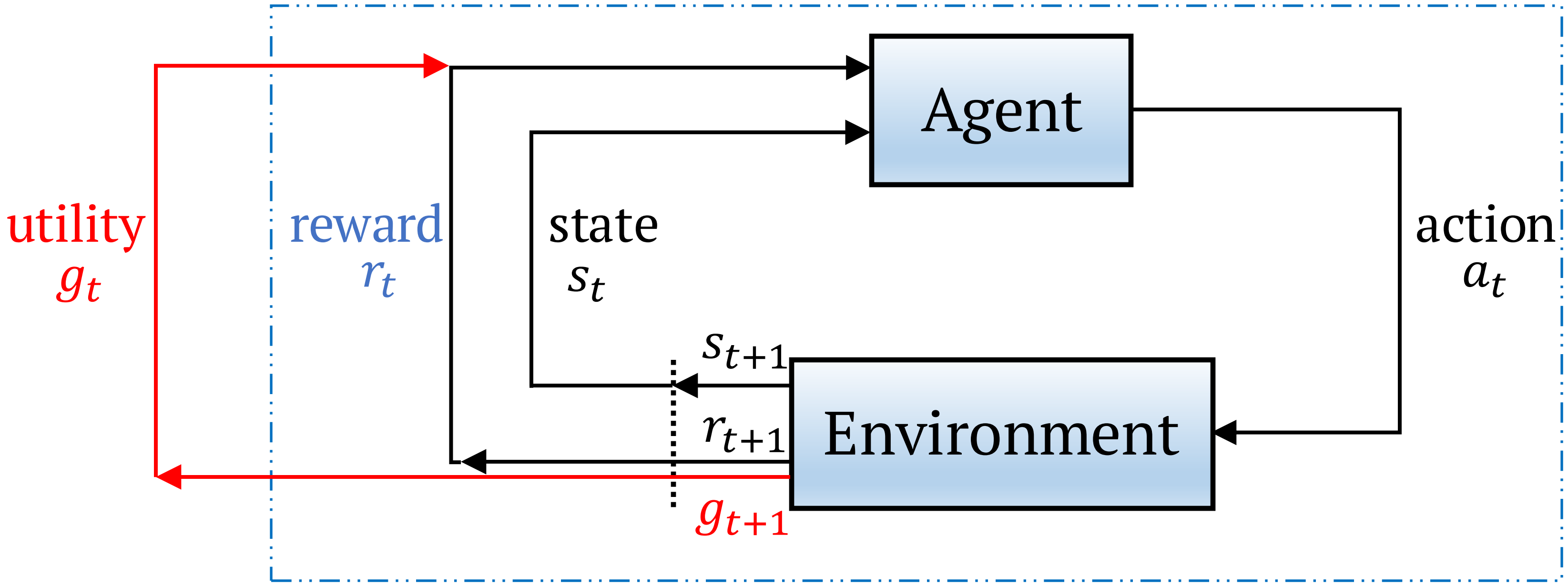
Dongsheng Ding, Kaiqing Zhang, Tamer Başar, Mihailo R. Jovanović

Thirty-fourth Conference on Neural Information Processing Systems, Dec 6th -12th , 2020

Constrained Reinforcement Learning



Constrained Reinforcement Learning



Constrained Reinforcement Learning

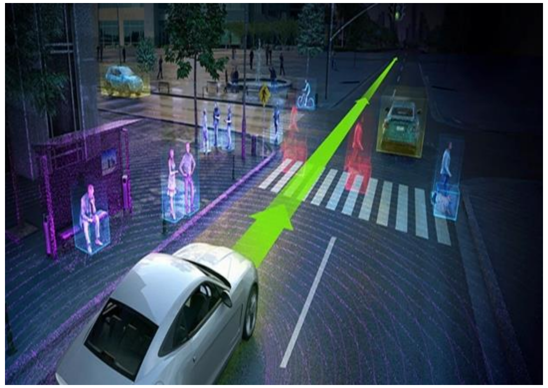
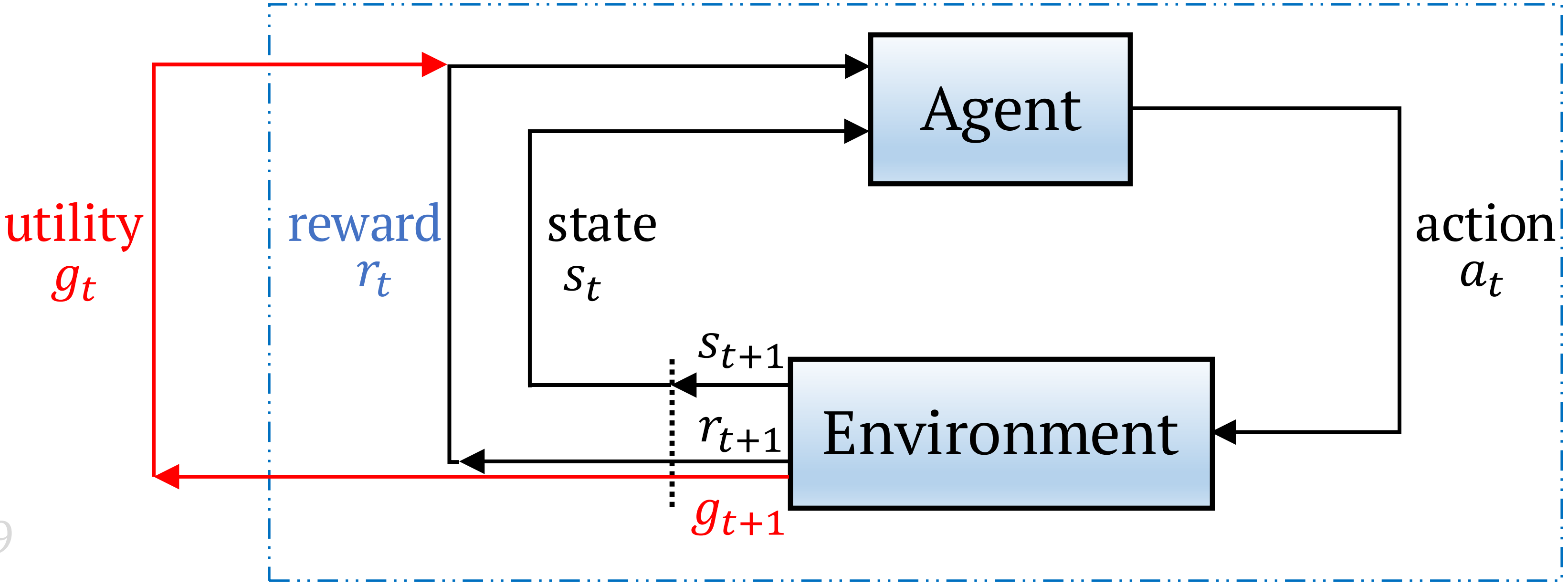
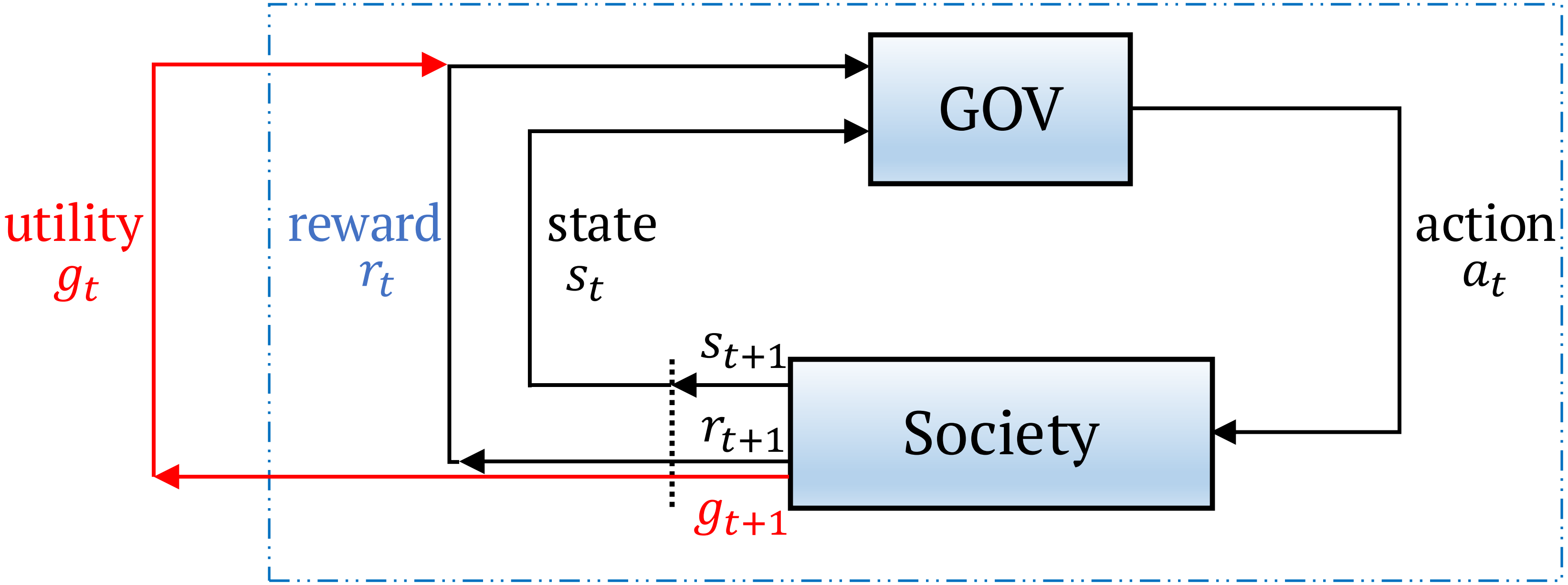


Figure: Choudhury, AIM ’19



Example: COVID-19 Pandemic

Figure: IHRB '20



Constrained Markov Decision Processes

$$\text{maximize} \quad V_r^\pi(\rho)$$

Reward maximization

$$\text{subject to} \quad V_g^\pi(\rho) \;\geq\; b$$

Utility constraint

$$\pi \in \Pi$$

- $V_r^\pi(\rho) = \mathrm{E}_{\pi,\, s_0 \sim \rho}\left[\, \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid \pi, s_0 \,\right]$
- $V_g^\pi(\rho) = \mathrm{E}_{\pi,\, s_0 \sim \rho}\left[\, \sum_{t=0}^{\infty} \gamma^t g(s_t, a_t) \mid \pi, s_0 \,\right]$
- $s_0 \sim \rho, a_t \sim \pi(\cdot \mid s_t), s_{t+1} \sim P(\cdot \mid s_t, a_t)$

Altman, '99

Direct Policy Search

- Lagrangian-Based Actor-Critic

  (Borkar, 2004), (Bhatnagar, Lakshmanan, 2004), (Chow, Ghavamzadeh, Janson, Pavone, 2017), (Tessler, Mankowitz, Mannor, 2019), (Spooner, Savani, 2020), et al.

- Constrained Policy Gradient Method

  (Uchibe, Doya, 2009), et al.

- Constrained Policy Optimization

  (Achiam, Held, Tamar, Abbeel, 2017), (Yang, Rosca, Narasimhan, Ramadge, 2020), (Liu, Ding, Liu, 2020), et al.

- Lagrangian-Based Policy Optimization

  (Liang, Que, Modiano, 2018), (Paternain, Chamon, Calvo-Fullanan, Ribeiro, 2019), (Stooke, Achiam, Abbeel, 2020), et al.

Natural Policy Gradient Primal-Dual Method

- Primal Update

$$\theta^{(t+1)} = \theta^{(t)} + \eta_1 F_\rho(\theta^{(t)})^\dagger \cdot \nabla_\theta V_L^{\theta^{(t)}, \lambda^{(t)}}(\rho)$$

- Dual Update

$$\lambda^{(t+1)} = \mathcal{P}\left[\lambda^{(t)} - \eta_2 \nabla_\theta V_L^{\theta^{(t)}, \lambda^{(t)}}(\rho))\right]$$

$$\text{maximize}_{\theta} \ \text{minimize}_{\lambda \geq 0} \ \underbrace{V_r^{\pi_\theta}(\rho) + \lambda(V_g^{\pi_\theta}(\rho) - b)}_{V_L^{\pi_\theta, \lambda}(\rho)}$$

Non-Asymptotic Convergence

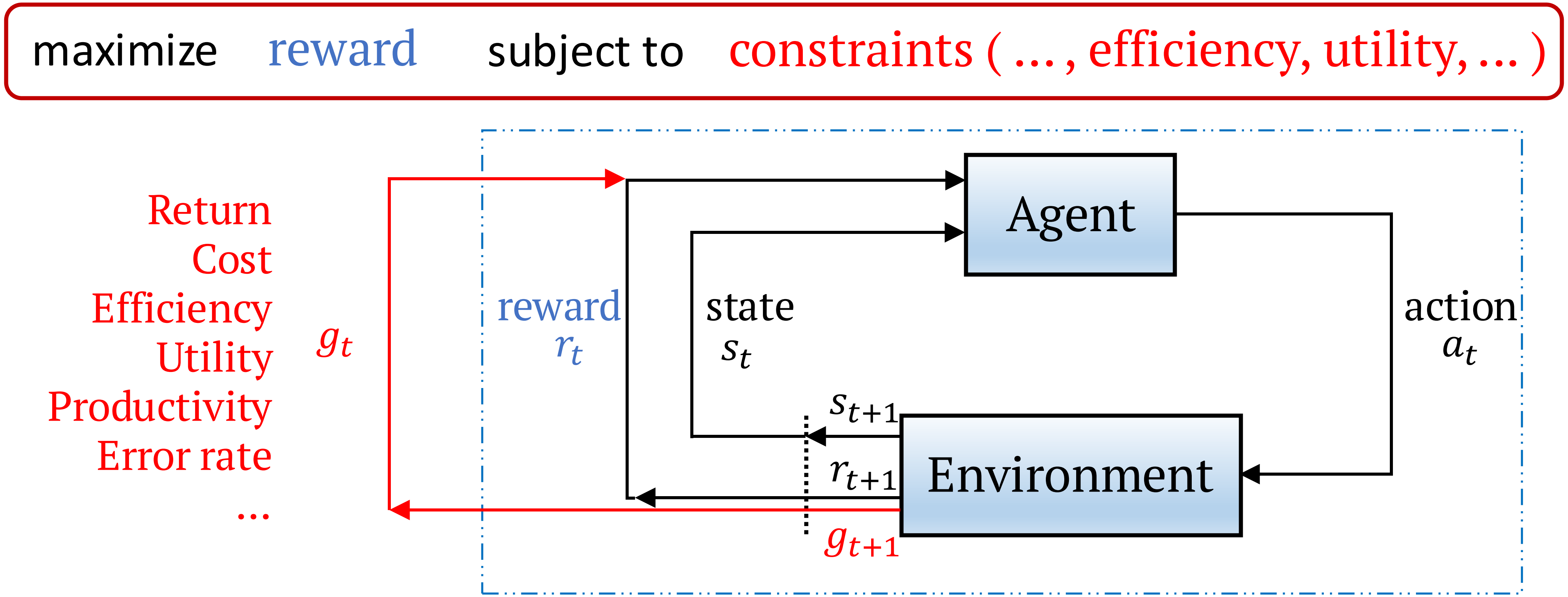
Policy Class	Optimality Gap	Constraint Violation
Softmax Policy	$O\left(\frac{1}{\sqrt{T}}\right)$	$O\left(\frac{1}{\sqrt{T}}\right)$
General Policy Parametrization	$O\left(\frac{1}{\sqrt{T}} + \sqrt{\epsilon}\right)$	$O\left(\frac{1}{T^{1/4}} + \left(\frac{\epsilon}{T}\right)^{1/4}\right)$

- T – the total number of gradient iterations
- ϵ – the function approximation error

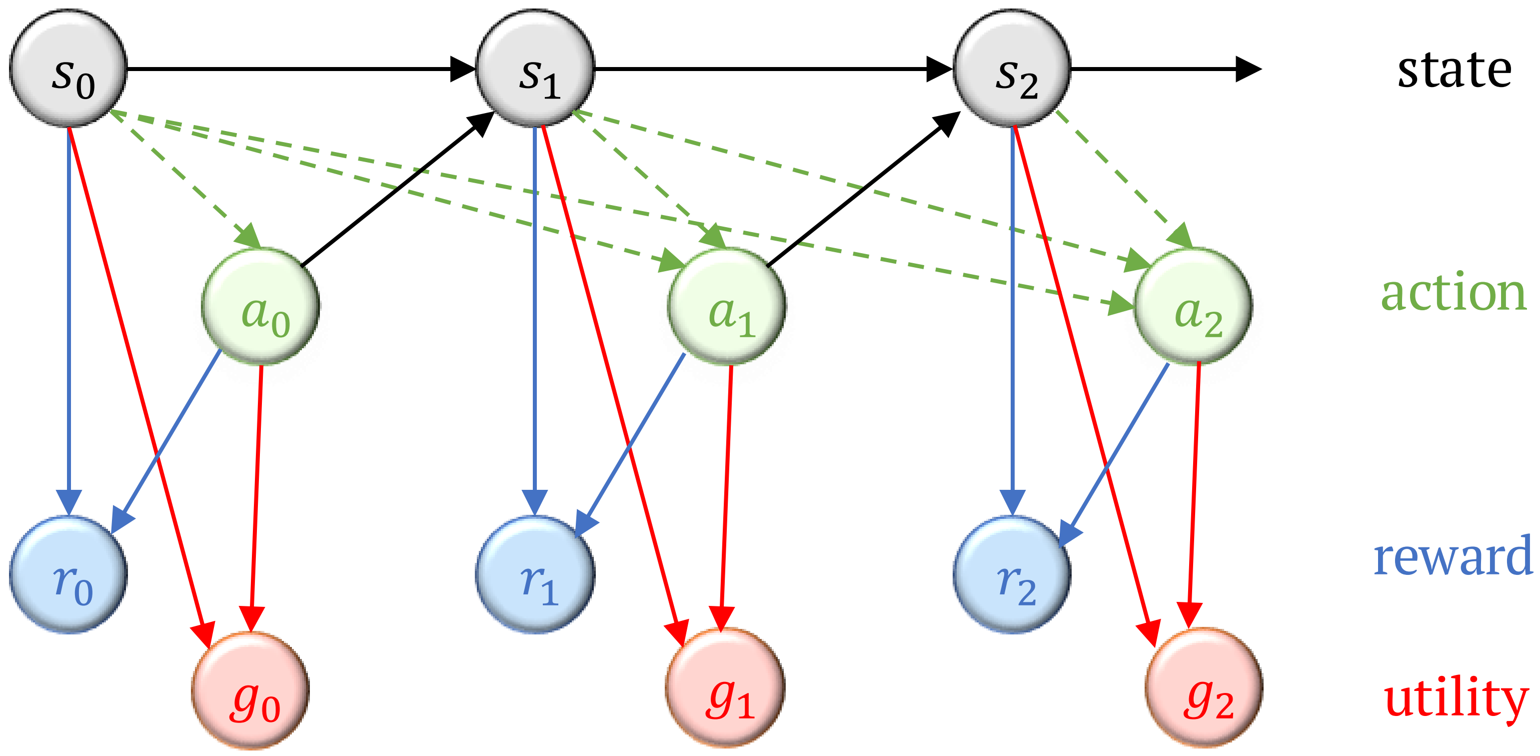
- O – has no dimension-dependence
- O – has only $\log |\mathcal{A}|$

Safe Reinforcement Learning

Figure: Cardinal, ExtremeTech '18



maximize reward subject to constraints (… , efficiency, utility, …)
Return
Cost
Efficiency
Utility
Productivity
Error rate
…
g_t
reward
r_t
state
s_t
Agent
action
a_t
s_{t+1}
r_{t+1}
g_{t+1}
Environment



s_0
s_1
s_2
state
a_0
a_1
a_2
action
r_0
r_1
r_2
reward
g_0
g_1
g_2
utility