# Reinforcement Learning Lab 5

## Policy gradient

Niccolò Turcato (niccolo.turcato@phd.unipd.it)

Department of Information Engineering - Università Degli studi di Padova

# Policy gradient for linear function approximation

Define:

$$h(s,a,\theta) = \theta^T x(s,a) \tag{1}$$

and consider Policy:

$$\pi(a|s,\theta) = \frac{e^{h(s,a,\theta)}}{\Sigma_b e^{h(s,b,\theta)}} \tag{2}$$

The gradient is defined as:

$$\nabla \hat{J}(\theta) = G_t \nabla_\theta ln(\pi(a|s,\theta)) \tag{3}$$

# Computation

$$\nabla_\theta ln(\pi(a|s,\theta)) = \nabla_\theta ln\left(\frac{e^{h(s,a,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}\right) = \nabla_\theta \left(ln(e^{h(s,a,\theta)}) - ln(\Sigma_b e^{h(s,b,\theta)})\right)$$

# Computation

$$
\begin{aligned}
\nabla_\theta ln(\pi(a|s,\theta)) &= \nabla_\theta ln\left(\frac{e^{h(s,a,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}\right) = \nabla_\theta \left(ln(e^{h(s,a,\theta)}) - ln(\Sigma_b e^{h(s,b,\theta)})\right) \\
&= \nabla_\theta h(s,a,\theta) - \nabla_\theta ln(\Sigma_b e^{h(s,b,\theta)}) = x(s,a) - \frac{\nabla_\theta \Sigma_b e^{h(s,b,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}
\end{aligned}
$$

# Computation

$$\nabla_\theta ln(\pi(a|s,\theta)) = \nabla_\theta ln\left(\frac{e^{h(s,a,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}\right) = \nabla_\theta\left(ln(e^{h(s,a,\theta)}) - ln(\Sigma_b e^{h(s,b,\theta)})\right)$$

$$= \nabla_\theta h(s,a,\theta) - \nabla_\theta ln(\Sigma_b e^{h(s,b,\theta)}) = x(s,a) - \frac{\nabla_\theta \Sigma_b e^{h(s,b,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}$$

$$= x(s,a) - \frac{\Sigma_b e^{h(s,b,\theta)}\nabla_\theta h(s,b,\theta)}{\Sigma_b e^{h(s,b,\theta)}} = x(s,a) - \frac{\Sigma_b e^{h(s,b,\theta)}x(s,b)}{\Sigma_b e^{h(s,b,\theta)}}$$

# Computation

$$
\begin{aligned}
\nabla_\theta ln(\pi(a|s,\theta)) &= \nabla_\theta ln\left(\frac{e^{h(s,a,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}\right) = \nabla_\theta\left(ln(e^{h(s,a,\theta)}) - ln(\Sigma_b e^{h(s,b,\theta)})\right) \\
&= \nabla_\theta h(s,a,\theta) - \nabla_\theta ln(\Sigma_b e^{h(s,b,\theta)}) = x(s,a) - \frac{\nabla_\theta \Sigma_b e^{h(s,b,\theta)}}{\Sigma_b e^{h(s,b,\theta)}} \\
&= x(s,a) - \frac{\Sigma_b e^{h(s,b,\theta)} \nabla_\theta h(s,b,\theta)}{\Sigma_b e^{h(s,b,\theta)}} = x(s,a) - \frac{\Sigma_b e^{h(s,b,\theta)} x(s,b)}{\Sigma_b e^{h(s,b,\theta)}} \\
&= x(s,a) - \Sigma_b\left(\frac{e^{h(s,b,\theta)}}{\Sigma_b e^{h(s,b,\theta)}}\right) x(s,b) = x(s,a) - \Sigma_b \pi(b|s,\theta) x(s,b)
\end{aligned}
$$