

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

The REINFORCE Algorithm

The REINFORCE Algorithm

Williams, 1992

REINFORCE is a Policy Gradient Algorithm

We assume a parameterized policy $\pi_{\Phi}(a|s)$.

$\pi_{\Phi}(a|s)$ is normalized while $Q_{\Phi}(s, a)$ is not.

Policy Gradient Theorem (Episodic Case)

$$\Phi^* = \operatorname{argmax}_{\Phi} E_{\pi_{\Phi}} R$$

$$\nabla_{\Phi} E_{\pi_{\Phi}} R = \sum_{s_0, a_0, s_1, a_1, \dots, s_T, a_T} \nabla_{\Phi} P(s_0, a_0, s_1, a_1, \dots, s_T, a_T) R$$

$$\begin{aligned} \nabla_{\Phi} P(\dots) R &= P(S_0) \nabla_{\Phi} \pi(a_0) P(s_1) \pi(a_1) \cdots P(s_T) \pi(a_T) R \\ &\quad + P(S_0) \pi(a_0) P(s_1) \nabla_{\Phi} \pi(a_1) \cdots P(s_T) \pi(a_T) R \\ &\quad \vdots \\ &\quad + P(S_0) \pi(a_0) P(s_1) \pi(a_1) \cdots P(s_T) \nabla_{\Phi} \pi(a_T) R \end{aligned}$$

$$= P(\dots) \left(\sum_t \frac{\nabla_{\Phi} \pi_{\Phi}(a_t)}{\pi_{\Phi}(a_t)} \right) R$$

Policy Gradient Theorem (Episodic Case)

$$\nabla_{\Phi} P(\dots)R = P(\dots) \left(\sum_t \frac{\nabla_{\Phi} \pi_{\Phi}(a_t|s_t)}{\pi_{\Phi}(a_t|s_t)} \right) R$$

$$\nabla_{\Phi} E_{\pi_{\Phi}} R = E_{\pi_{\Phi}} \left(\sum_t \nabla_{\Phi} \ln \pi_{\Phi}(a_t|s_t) \right) R$$

Policy Gradient Theorem

$$\begin{aligned} & \nabla_{\Phi} E_{\pi_{\Phi}} R \\ &= E_{\pi_{\Phi}} \left(\sum_t \nabla_{\Phi} \ln \pi_{\Phi}(a_t|s_t) \right) R \\ &= E_{\pi_{\Phi}} \left(\sum_t \nabla_{\Phi} \ln \pi_{\Phi}(a_t|s_t) \right) \left(\sum_t r_t \right) \\ &= E_{\pi_{\Phi}} \sum_{t,t'} \nabla_{\Phi} \ln \pi_{\Phi}(a_t|s_t) r_{t'} \end{aligned}$$

Policy Gradient Theorem

$$\nabla_{\Phi} E_{\pi_{\Phi}} R = \sum_{t,t'} E_{s_t, a_t, r_{t'}} \nabla_{\Phi} \ln \pi_{\Phi}(a_t | s_t) r_{t'}$$

For $t' < t$ we have

$$\begin{aligned} E_{r_{t'}, s_t, a_t} r_{t'} \nabla_{\Phi} \ln \pi_{\Phi}(a_t | s_t) &= E_{r_{t'}, s_t} r_{t'} \sum_{a_t} \pi_{\Phi}(a_t | s_t) \nabla_{\Phi} \ln \pi_{\Phi}(a_t | s_t) \\ &= E_{r_{t'}, s_t} r_{t'} \sum_{a_t} \nabla_{\Phi} \pi_{\Phi}(a_t | s_t) \\ &= E_{r_{t'}, s_t} r_{t'} \nabla_{\Phi} \sum_{a_t} \pi_{\Phi}(a_t | s_t) \\ &= 0 \end{aligned}$$

REINFORCE

$$\nabla_{\Phi} E_{\pi_{\Phi}} R = E_{\pi_{\Phi}} \sum_{t, t' \geq t} (\nabla_{\Phi} \ln \pi_{\Phi}(a_t | s_t)) r_{t'}$$

Sampling runs and computing the above sum over t and t' is Williams' REINFORCE algorithm.

Optimizing Discrete Decisions with Non-Differentiable Loss

The REINFORCE algorithm is used generally for non-differentiable loss functions.

For example error rate and BLEU score are non-differentiable — they are defined on the result of discrete decisions.

$$\Phi^* = \operatorname{argmax}_{\Phi} E_{w_1, \dots, w_n \sim P_{\Phi}} \text{BLEU}$$

END