Reinforcement Learning May 15, 2019

Recitation 9

Lecturer:Yishay Mansour *TA: Lee Cohen*

Definition 1.

REINFORCE

Input:a differentiable policy parameterization $\pi(a|s,\theta)$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$
repeat
  Generate an episode $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$.
  for $t = 0, 1, ..., T-1$ **do**
    $G \leftarrow$ return from step t
    $\theta \leftarrow \theta + \alpha G_t \nabla_\theta ln\pi(A_t|S_t,\theta)$
until Forever

Definition 2.

REINFORCE with Baseline

Input:a differentiable policy parameterization $\pi(a|s,\theta)$
Input:a differentiable state-value parameterization $\hat{v}(s;w)$
Parameters: step sizes $\alpha^w > 0, \alpha^\theta > 0$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state-value weights $w \in \mathbb{R}^d$
repeat
  Generate an episode $S_0, A_0, R_1, ..., S_{T-1}, A_{T-1}, R_T$ following $\pi(\cdot|\cdot,\theta)$.
  for $t = 0, 1, ..., T-1$ **do**
    $G_t \leftarrow$ return from step t
    $\delta \leftarrow G_t - \hat{v}(S_t, w)$
    $w \leftarrow w + \alpha^w \delta \nabla_w \hat{v}(S_t; w)$
    $\theta \leftarrow \theta + \alpha^\theta \gamma^t \delta \nabla_\theta ln\pi(A_t|S_t,\theta)$
until Forever

Definition 3.

Actor-Critic

Input:a differentiable policy parameterization $\pi(a|s,\theta)$ - actor
Input:a differentiable state-value parameterization $\hat{Q}(s;w)$ - critic
Parameters: step sizes $\alpha^w > 0, \alpha^\theta > 0$

Initialize policy parameter $\theta \in \mathbb{R}^{d'}$ and state-value weights $w \in \mathbb{R}^d$
repeat
    Initialize S (first state of episode).
    while S is not terminal **do**
        $A \sim \pi(\cdot|S,\theta)$
        Take action A, observe S',R
        $A' \sim \pi(\cdot|S',\theta)$
        $\delta \leftarrow R + \gamma\hat{Q}(S',A';w) - \hat{Q}(S,A;w)$ (if S' is terminal, then $\hat{Q}(S',A';w) = 0$)
        $w \leftarrow w + \alpha^w\delta\nabla_w\hat{Q}(S,A;w)$
        $\theta \leftarrow \theta + \alpha^\theta\gamma^t\delta\nabla_\theta ln\pi(A_t|S_t,\theta)$
        $S \leftarrow S'$
until Forever

Discrete Action Space

Exercise 1.
A Bernoulli-logistic unit is a stochastic neuron-like unit used in some artificial neural networks. Its input at time t is a feature vector $x(S_t)$; its output, A_t, is a random variable having two values, 0 and 1, with $Pr\{A_t = 1\} = P_t$ and $Pr\{A_t = 0\} = 1 - P_t$ (the Bernoulli distribution).

Let $h(s, 0, \theta)$ and $h(s, 1, \theta)$ be the preferences in state s for the unit's two actions given policy parameter θ. Assume that the difference between the preferences is given by a weighted sum of the unit's input vector, that is, assume that $h(s, 1, \theta) - h(s, 0, \theta) = \theta^T x(s)$, where θ is the unit's weight vector.

1. Show that if the exponential softmax distribution is used to convert preferences to policies, then $P_t = \pi(1|S_t, \theta_t) = 1/(1 + exp(-\theta_t^T x(S_t))$ (the logistic function).
2. What is the Monte-Carlo REINFORCE update of θ_t to θ_{t+1} upon receipt of return G_t?
3. Express the eligibility $\nabla_\theta ln(\pi(a|s, \theta))$ for a Bernoulli-logistic unit, in terms of a,$x(s)$, and $\pi(a|s, \theta)$ by calculating the gradient.

Hint: separately for each action compute the derivative of the logorithm first with respect to $P_t = \pi(a|s, \theta_t)$, combine the two results into one expression that depends on a and Pt, and then use the chain rule, noting that the derivative of the logistic function $f(x)$ is $f(x)(1 - f(x))$.

Solution 1. Recall, the softmax activation: Given a sample vector x, the output of softmax activation is a distribution vector z:

$$z_i = \frac{e^{x_i}}{\sum_j e^{x_j}}$$

1.

$$P_t = \pi(1|s, \theta_t) = \frac{e^{h(s,1,\theta)}}{e^{h(s,1,\theta)} + e^{h(s,0,\theta)}}$$

Divide by $e^{h(s,1,\theta)}$ both nominator and denominator:

$$P_t = \frac{1}{1 + e^{h(s,0,\theta) - h(s,1,\theta)}} = \frac{1}{1 + e^{-\theta^T x(s)}}$$

Note that this is the logistic function, which derivative is: $f'(x) = f(x)(1 - f(x))$

2. $\theta \leftarrow \theta + \alpha G \nabla_\theta ln\pi(A_t|S_t, \theta)$

3. To simplify the derivative we define $g(s;\theta) = \theta^T x(s)$. It follows that

$$P_t = \frac{1}{1 + e^{-g(s;\theta)}}$$

and the derivative of P_t with respect to $g(s;\theta)$ is:

$$P_t' = P_t(1 - P_t)$$

Now, we will derive separately for each action using $g(s;\theta)$ and then combine the two results:

$$\nabla_g ln(\pi(1|s,\theta)) = \nabla_g ln(P_t) = \frac{P_t(1-P_t)}{P_t} = 1 - P_t$$
$$\nabla_g ln(\pi(0|s,\theta)) = \nabla_g ln(1 - P_t) = -\frac{P_t(1-P_t)}{1-P_t} = -P_t$$

Now, we can combine the two based on action a and derive based on θ:

$$\nabla_\theta ln(\pi(a|s,\theta)) = [a(1-P_t) - (1-a)P_t] \cdot \nabla_\theta(g) = [a(1-P_t) - (1-a)P_t]x(s).$$

Continuous Action Space

Exercise 2.
Policy-based methods offer practical ways of dealing with large actions spaces, even continuous spaces with an infinite number of actions. Instead of computing learned probabilities for each of the many actions, we instead learn statistics of the probability distribution. For example, the action set might be the real numbers, with actions chosen from a normal (Gaussian) distribution. The probability density function for the normal distribution is conventionally written:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

where μ and σ here are the mean and standard deviation of the normal distribution. The probability density functions for several different means and standard deviations are shown in Fig. 1. The value p(x) is the density of the probability at x, not the probability. It can be greater than 1; it is the total area under $p(x)$ that must sum to 1. In general, one can take the integral under $p(x)$ for any range of x values to get the probability of x falling within that range.

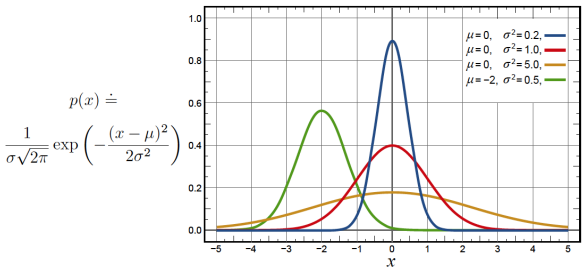

Figure 1: The probability density function of the normal distribution for different means and variances.

To produce a policy parameterization, the policy can be defined as the normal probability density over a real-valued scalar action, with mean and standard deviation given by parametric function approximators that depend on the state. That is,

$$\pi(a|s,\theta) = \frac{1}{\sigma(s,\theta)\sqrt{2\pi}} exp(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2})$$

The mean can be approximated as a linear function. The standard deviation must always be positive and is better approximated as the exponential of a linear function. Thus

$$\mu = \theta_\mu^T x(s), \quad \sigma = exp(\theta_\sigma^T x(s))$$

Derive the weight updates for the Gaussian policy parameterization.

Solution 2. For this we divide the policy's parameter vector into two parts, $[\theta_\mu; \theta_\sigma]$, one part to be used for the approximation of the mean and one part for the approximation of the standard deviation.

The gradient of the expectation:

$$\nabla_{\theta_\mu} ln(\pi(a|s,\theta)) = \nabla_{\theta_\mu}(-\frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}) = \frac{-1}{2\sigma(s,\theta)^2}(2(a-\mu(s,\theta)) \cdot (-x(s)) = \frac{a-\mu(s,\theta)}{\sigma(s,\theta)^2}x(s)$$

The gradient of the standard deviation:

$$\nabla_{\theta_\sigma} ln(\pi(a|s,\theta)) = \nabla_{\theta_\sigma}[ln(\frac{1}{\sqrt{2\pi}}) + ln(\frac{1}{\sigma(s,\theta)}) - \frac{(a-\mu(s,\theta))^2}{2\sigma(s,\theta)^2}] =$$
$$[\sigma(s,\theta)\frac{-1}{\sigma(s,\theta)^2} - \frac{-4\sigma(s,\theta)(a-\mu(s,\theta))^2}{4\sigma(s,\theta)^4}]\nabla_{\theta_\sigma}\sigma(s,\theta) =$$
$$[\sigma(s,\theta)\frac{-1}{\sigma(s,\theta)^2} - \frac{-4\sigma(s,\theta)(a-\mu(s,\theta))^2}{4\sigma(s,\theta)^4}]\sigma(s,\theta)x(s) =$$
$$[-1 + \frac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^2}]x(s) =$$
$$[\frac{(a-\mu(s,\theta))^2}{\sigma(s,\theta)^2} - 1]x(s)$$