NX-414
A. Mathis
M. Schrimpf

Revision - Part 2

Reinforcement learning

2023, Exercise 3.1

Consider the definitions of v function and q function:

$$v(s_t) = \mathbb{E}\left[G_t | S_t = s_t\right] \qquad \text{and} \qquad q(s_t, a_t) = \mathbb{E}\left[G_t | S_t = s_t, A_t = a_t\right], \tag{1}$$

where s_t and a_t are the state and the action at step t of a Markov decision process, S_t and A_t are the corresponding random variables, γ is the discount factor and $G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k}$ is the discounted cumulative reward (R_t is the random variable representing the reward obtained at step t). Provide a proof of the following relations linking the q function and the v function:

$$v(s_t) = \int_{\mathcal{A}} \pi(a_t | S_t = s_t) q(s_t, a_t)\, da_t \tag{2}$$

and

$$q(s_t, a_t) = \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t | S_t = s_t, A_t = a_t)[r_t + \gamma v(s_{t+1})]\, ds_{t+1}\, dr_t, \tag{3}$$

where $\pi(a_t | S_t = s_t)$ is the action distribution defined by the policy and $p(s_{t+1}, r_t | S_t = s_t, A_t = a_t)$ is the transition probability distribution, defining the probability of reaching the state s_{t+1} choosing the action a_t in the state s_t, and of obtaining a reward r_t in the process. Here $\mathcal{A}$, $\mathcal{S}$, and $\mathcal{R}$ are the action, state and reward spaces, respectively.[a]. In your calculations, feel free to use the notation abuse for the conditioning to the observed values of states and actions, so that, e.g., $\mathbb{E}\left[G_t | S_t = s_t\right]$ can be written $\mathbb{E}\left[G_t | s_t\right]$ and $p(s_{t+1}, r_t | S_t = s_t, A_t = a_t)$ can be written $p(s_{t+1}, r_t | s_t, a_t)$.
Note: Exercise 3.2 can be done independently of 3.1.

[a] Remember the law of total expectations for conditional expected values:
$\mathbb{E}\left[A | B = b\right] = \int_{\mathcal{C}} p(c | B = b) \mathbb{E}\left[A | B = b, C = c\right]$

Solution: The first result comes from a direct application of total expectations:

$$v(s_t) = \mathbb{E}\left[G_t | s_t\right] = \int_{\mathcal{A}} \pi(a_t | s_t) \mathbb{E}\left[G_t | s_t, a_t\right]\, da_t = \int_{\mathcal{A}} \pi(a_t | s_t) q(s_t, a_t)\, da_t. \tag{4}$$

The second result can be obtained in a similar way:

$$\begin{aligned} q(s_t, a_t) &= \mathbb{E}\left[G_t | s_t, a_t\right] = \mathbb{E}\left[R_t + \gamma G_{t+1} | s_t, a_t\right] \\ &= \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t | s_t, a_t)[r_t + \gamma \mathbb{E}\left[G_{t+1} | s_t, a_t, s_{t+1}\right]]\, dr_t\, ds_{t+1} \\ &= \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t | s_t, a_t)[r_t + \gamma \mathbb{E}\left[G_{t+1} | s_{t+1}\right]]\, dr_t\, ds_{t+1} \\ &= \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t | s_t, a_t)[r_t + \gamma v(s_{t+1})]\, dr_t\, ds_{t+1}, \end{aligned} \tag{5}$$

where we have used: definition of G_t, definition of expected value of R_t, law of total expectations, Markov property and definition of $v(s_t)$.

2023, Exercise 3.2

Use Eq. 2 and Eq. 3 to derive the Bellman equations for the v function and for the q function. Discuss the meaning of these equations (you are encouraged to draw a diagram to improve the clarity of your explanation).

Solution: The Bellman equations for v and q can be obtained by simply combining the expressions obtained from the previous exercise:

$$v(s_t) = \int_{\mathcal{A}} \pi(a_t|s_t) \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t|s_t, a_t)[r_t + \gamma v(s_{t+1})]\, dr_t\, ds_{t+1}\, da_t \tag{6}$$

and

$$q(s_t, a_t) = \int_{\mathcal{S}} \int_{\mathcal{R}} p(s_{t+1}, r_t|s_t, a_t)[r_t + \gamma \int_{\mathcal{A}} \pi(a_{t+1}|s_{t+1}) q(s_{t+1}, a_{t+1})\, da_{t+1}]\, dr_t\, ds_{t+1} \tag{7}$$

2024, Exercise 3.1

In Reinforcement Learning, policy gradient methods learn a *parametrized policy* that can select actions without considering a value function. We denote the policy parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$, such that $\pi(a|s, \boldsymbol{\theta}) = \Pr\{A_t = a|S_t = s, \boldsymbol{\theta}_t = \boldsymbol{\theta}\}$ is the probability that action a is taken at time t given that the environment is in state s at time t with parameter $\boldsymbol{\theta}$.
If the action space is discrete and not too large, then a natural parametrization is to form parametrized numerical preferences $h(s, a, \boldsymbol{\theta}) \in \mathbb{R}$ for each state-action pair. The probability of an action being selected is then given according to an exponential soft-max distribution:

$$\pi(a|s, \boldsymbol{\theta}) \doteq \frac{e^{h(s,a,\boldsymbol{\theta})}}{\sum_b e^{h(s,b,\boldsymbol{\theta})}}$$

Suppose that the action space is the binary set $\{0, 1\}$, and let $h(s, 0, \boldsymbol{\theta})$ and $h(s, 1, \boldsymbol{\theta})$ be the preferences in state s for the two actions given policy parameter $\boldsymbol{\theta}$. Assume that a state is defined by a feature vector $\mathbf{x}(S_t)$, and that we can express the difference between the action preferences as:

$$h(s, 1, \boldsymbol{\theta}) - h(s, 0, \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}(s)$$

Show that if the exponential soft-max distribution is used to convert action preferences to policies, then the probability of taking action $a = 1$ is given by:

$$\pi(a = 1|s, \boldsymbol{\theta}) = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}.$$

This is just a simple combination of the two given expressions to arrive at the desired formula. We apply the definition of the soft-max policy to the binary action space to get:

$$\pi(a = 1|s, \boldsymbol{\theta}) = \frac{e^{h(s,1,\boldsymbol{\theta})}}{e^{h(s,1,\boldsymbol{\theta})} + e^{h(s,0,\boldsymbol{\theta})}} = \frac{1}{1 + e^{h(s,0,\boldsymbol{\theta}) - h(s,1,\boldsymbol{\theta})}} = \frac{1}{1 + e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}.$$

2024, Exercise 3.2

Express the eligibility $\nabla_\theta \log \pi(a|s, \boldsymbol{\theta})$ for the above parametrization, in terms of a, $\mathbf{x}(s)$, and $\pi(a|s, \boldsymbol{\theta})$.

We begin by expressing the policy in closed form:

$$\pi(a|s,\boldsymbol{\theta}) = \left(\frac{1}{1+e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}\right)^a \left(1-\frac{1}{1+e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}\right)^{1-a} = \left(\frac{1}{1+e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}\right)^a \left(\frac{1}{1+e^{\boldsymbol{\theta}^\top \mathbf{x}(s)}}\right)^{1-a}$$

Taking the logarithm yields:

$$\log \pi(a|s,\boldsymbol{\theta}) = -a \log\left(1+e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}\right) + (a-1)\log\left(1+e^{\boldsymbol{\theta}^\top \mathbf{x}(s)}\right).$$

The derivative is then computed as:

$$\nabla_\theta \log \pi(a|s,\boldsymbol{\theta}) = a\mathbf{x}(s)\frac{e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}}{1+e^{-\boldsymbol{\theta}^\top \mathbf{x}(s)}} + (a-1)\mathbf{x}(s)\frac{e^{\boldsymbol{\theta}^\top \mathbf{x}(s)}}{1+e^{\boldsymbol{\theta}^\top \mathbf{x}(s)}} = a\mathbf{x}(s)\pi(0|s,\boldsymbol{\theta}) + (a-1)\mathbf{x}(s)\pi(1|s,\boldsymbol{\theta}),$$

which we can finally simplify as

$$\nabla_\theta \log \pi(a|s,\boldsymbol{\theta}) = (-1)^{1-a}\mathbf{x}(s)\left(1-\pi(a|s,\boldsymbol{\theta})\right)$$

2024, Exercise 3.3

Propose an update rule for the parameter $\boldsymbol{\theta}_t$ upon receipt of return G_t. (Note that $\mathbb{E}_\pi[G_t|S_t, A_t] = q_\pi(S_t, A_t)$.)

Recall equation (16) from the solutions of Week 13, which describes such an update rule:

$$\Theta_{i+1} = \Theta_i + \alpha q(s_t, a_t)\nabla_\Theta \log \pi(a_t|s_t),$$

where α is the step size. We mentioned that, in practice, there are several methods for estimating $q(s_t, a_t)$, and the hint points us to the Monte-Carlo REINFORCE algorithm:

$$\Theta_{i+1} = \Theta_i + \alpha G_t \nabla_\Theta \log \pi(a_t|s_t),$$

which works because G_t is an unbiased estimate of $q(s_t, a_t)$.

Attention

2023, Exercise 2.1

The fundamental building block of transformers is the rescaled dot-product attention, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{C}\right)V, \tag{8}$$

where $Q, K \in \mathbb{R}^{d_i \times d_k}$ are the query and key matrices, $V \in \mathbb{R}^{d_i \times d_v}$ is the value matrix and $C \in \mathbb{R}^+$ is a positive rescaling constant[a]. The softmax operation is computed along each row of $\frac{QK^\top}{C}$.
Assuming that the elements of Q and K are independent and distributed as standard Gaussian distributions, what is the mean and the variance of each element of $\frac{QK^\top}{C}$?

[a] Remember: $[\text{softmax}(\mathbf{x})]_i = \frac{\exp(x_i)}{\sum_j \exp(x_j)}$

Solution: We write the expression of the element (i, j) of the attention matrix (before softmax):

$$\left(\frac{QK^\top}{C}\right)_{i,j} = \frac{1}{C}\sum_{t=1}^{d_k} q_{i,t}k_{j,t}. \tag{9}$$

We can then compute the mean:

$$\mathbb{E}\left[\left(\frac{QK^\top}{C}\right)_{i,j}\right] = \frac{1}{C}\sum_{t=1}^{d_k} \mathbb{E}[q_{i,t}]\mathbb{E}[k_{j,t}] = 0 \tag{10}$$

where we have used that all the variables are independent standard Gaussians. Similarly for the variance:

$$\mathbb{V}\left[\left(\frac{QK^\top}{C}\right)_{i,j}\right] = \frac{1}{C^2}\sum_{t=1}^{d_k} \mathbb{V}[q_{i,t}]\mathbb{V}[k_{j,t}] = \frac{d_k}{C^2}, \tag{11}$$

as all the elements of the sum are 1.

2023, Exercise 2.2

For which value of C is the variance of each element of $\frac{QK^\top}{C}$ equal to 1?

Solution: From the previous solution it immediately follows that, for $C = \sqrt{d_k}$, the variance is 1.

2023, Exercise 2.3

Compute $\frac{\partial}{\partial q_{1,1}}\left[\text{softmax}\left(\frac{QK^\top}{C}\right)\right]_{1,1}$, i.e., the partial derivative of $\left[\text{softmax}\left(\frac{QK^\top}{C}\right)\right]_{1,1}$ (i.e., the first element of the first row of the attention matrix) with respect to $q_{1,1}$ (i.e., the first element of the first row of the matrix Q) in the case where $d_i = 2$ (i.e., the matrices Q, K and V have 2 rows). Express the result as a function of C, $k_{1,1}$ and $k_{1,2}$ (i.e., the first elements of the first and the second rows of the matrix K) and of $T = \mathbf{q}_1^\top(\mathbf{k}_2 - \mathbf{k}_1)$. With $\mathbf{q}_1^\top$ and $\mathbf{q}_2^\top$ we refer to the two rows of the matrix Q, while with $\mathbf{k}_1^\top$ we refer to the first row of the matrix K.
*N*ote: this can be done fully independent of Exercise 2.1.

Solution:

$$\begin{aligned}\frac{\partial}{\partial q_{1,1}}\frac{\exp\left(\frac{\mathbf{q}_1^\top\mathbf{k}_1}{C}\right)}{\exp\left(\frac{\mathbf{q}_1^\top\mathbf{k}_1}{C}\right)+\exp\left(\frac{\mathbf{q}_1^\top\mathbf{k}_2}{C}\right)} &= \frac{\partial}{\partial q_{1,1}}\left(1+\exp\left(\frac{\mathbf{q}_1^\top(\mathbf{k}_2-\mathbf{k}_1)}{C}\right)\right)^{-1} \\ &= -\frac{1}{C}\left(1+\exp\left(\frac{\mathbf{q}_1^\top(\mathbf{k}_2-\mathbf{k}_1)}{C}\right)\right)^{-2}\exp\left(\frac{\mathbf{q}_1^\top(\mathbf{k}_2-\mathbf{k}_1)}{C}\right)(k_{2,1}-k_{1,1}) \\ &= \frac{k_{1,1}-k_{2,1}}{C}\frac{\exp\left(\frac{T}{C}\right)}{\left(1+\exp\left(\frac{T}{C}\right)\right)^2}\end{aligned} \tag{12}$$

2023, Exercise 2.4

Compute

$$\lim_{T\to\infty}\frac{\partial}{\partial q_{1,1}}\left[\text{softmax}\left(\frac{QK^\top}{C}\right)\right]_{1,1} \tag{13}$$

and

$$\mathbb{V}\left[\frac{T}{C}\right], \tag{14}$$

(i.e., the variance of T/C), under the assumption that all the elements of Q and K are distributed as independent standard Gaussian distributions.

Solution:

$$\lim_{T\to\infty}\frac{\partial}{\partial q_{1,1}}\left[\text{softmax}\left(\frac{QK^\top}{C}\right)\right]_{1,1} = \lim_{T\to\infty}\frac{k_{1,1}-k_{2,1}}{C}\frac{\exp\left(\frac{T}{C}\right)}{\left(1+\exp\left(\frac{T}{C}\right)\right)^2} = \lim_{T\to\infty}\frac{k_{1,1}-k_{2,1}}{C}\exp\left(-\frac{T}{C}\right) = 0 \tag{15}$$

$$\begin{aligned}\mathbb{V}\left[\frac{T}{C}\right] = \frac{1}{C^2}\mathbb{V}\left[\mathbf{q}_1^\top(\mathbf{k}_2-\mathbf{k}_1)\right] = \frac{1}{C^2}\mathbb{V}\left[\sum_{i=1}^{d_k}q_{1,i}(k_{2,i}-k_{1,i})\right] = \\ \frac{1}{C^2}\sum_{i=1}^{d_k}\mathbb{V}\left[q_{1,i}\right]\mathbb{V}\left[k_{2,i}-k_{1,i}\right] = \frac{2d_k}{C^2}\end{aligned} \tag{16}$$

2024, Exercise 2.5

Why might rescaling the product in the dot-product attention be important when backpropagating gradients (especially for large values of d_k)? Is the value of C found in Exercise 2.2 a good choice? Use the results found in the previous exercises to motivate your answers.

Solution: Because for large values of the attention matrix the gradient is very small (it goes to 0 exponentially). Therefore, we can expect problems backpropagating the gradients. Keeping the variance of the elements of the attention matrix under control can largely mitigate this problem. The rescaling factor $\sqrt{d_k}$ keeps the size of the elements of the attention matrix in a reasonable range (and thus their gradient).