Exercises - Policy Gradient Methods

Reinforcement Learning

1. This problem is the same as Exercise 13.3 and Exercise 13.4 in the textbook.
   (a) Consider a discrete action space $\mathcal{A}$ and the soft-max policy parametrization

$$\pi(a|s,\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,a))}{\sum_{b\in\mathcal{A}} \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}$$

   where $\boldsymbol{x}(s,a)$ are some features. Show that

$$\nabla \ln \pi(a|s,\boldsymbol{\theta}) = \boldsymbol{x}(s,a) - \sum_{b\in\mathcal{A}} \pi(b|s,\boldsymbol{\theta})\boldsymbol{x}(s,b)$$

   (b) Consider a continuous action space $\mathcal{A}$ where each $a \in \mathcal{A}$ is a scalar, and the Gaussian policy parametrization

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

   Let

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\theta}_\mu \\ \boldsymbol{\theta}_\sigma \end{bmatrix}$$

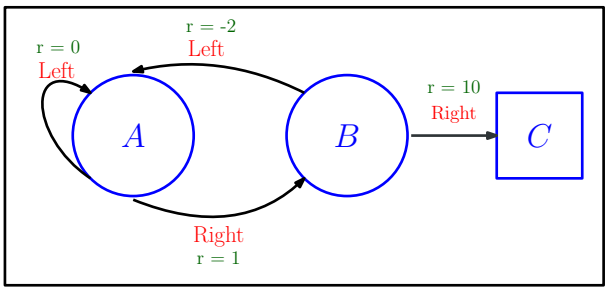
   and

$$\mu(s,\boldsymbol{\theta}) = \boldsymbol{\theta}_\mu^\top \boldsymbol{x}_\mu(s), \quad \sigma(s,\boldsymbol{\theta}) = \exp(\boldsymbol{\theta}_\sigma^\top \boldsymbol{x}_\sigma(s)).$$

   where $\boldsymbol{x}_\mu(s)$ and $\boldsymbol{x}_\theta(s)$ are features. Compute $\nabla \ln \pi(a|s,\boldsymbol{\theta})$.
   Remark: Note that with our parameter vector we have

$$\nabla f(\boldsymbol{\theta}) = \begin{bmatrix} \nabla_{\theta_\mu} f(\boldsymbol{\theta}) \\ \nabla_{\theta_\sigma} f(\boldsymbol{\theta}) \end{bmatrix}.$$

   where the subscript on ∇ indicates what we take the gradient with respect to. That is, you can compute the gradient with respect to $\boldsymbol{\theta}_\mu$ and $\boldsymbol{\theta}_\sigma$ separately.



Figur 2.1: A Markov Decision Process

2. Consider the MDP in Figure 2. The state space is $\mathcal{S} = \{A, B, C\}$, where C is a terminating state. The action space is $\mathcal{A} = \{\texttt{Left}, \texttt{Right}\}$. In this problem we will use a soft-max (see Problem 1 or

the textbook) policy with feature vectors given by

$$\boldsymbol{x}(A,\texttt{Left}) = \begin{bmatrix}1\\0\\0\\0\\0\\0\end{bmatrix}, \qquad \boldsymbol{x}(B,\texttt{Left}) = \begin{bmatrix}0\\1\\0\\0\\0\\0\end{bmatrix}, \qquad \boldsymbol{x}(C,\texttt{Left}) = \begin{bmatrix}0\\0\\1\\0\\0\\0\end{bmatrix}$$

$$\boldsymbol{x}(A,\texttt{Right}) = \begin{bmatrix}0\\0\\0\\1\\0\\0\end{bmatrix}, \qquad \boldsymbol{x}(B,\texttt{Right}) = \begin{bmatrix}0\\0\\0\\0\\1\\0\end{bmatrix}, \qquad \boldsymbol{x}(C,\texttt{Rigth}) = \begin{bmatrix}0\\0\\0\\0\\0\\1\end{bmatrix}$$

The parameter vector is initialized to

$$\boldsymbol{\theta}_0 = \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}$$

We consider an undiscounted setting, so $\gamma = 1$.

(a) With the initial parameter vector what give the probability of choosing `Left` vs `Right` when the agent in A and B?

(b) The agent used the initial policy in one episode resulting in the trajectory

$$S_0 = A, A_0 = \texttt{Right}, R_1 = 1, S_1 = B, A_1 = \texttt{Left}, R_2 = -2,$$
$$S_2 = A, A_2 = \texttt{Right}, R_3 = 1, S_3 = B, A_3 = \texttt{Right}, R_4 = 10, S_4 = C.$$

Use the REINFORCE to update the parameters using the first transition (S_0, A_0, R_1). You can assume that the learning rate is $\alpha = 0.1$.

(c) With the parameter vector obtained after this first update, what probability do the resulting policy give to each action in state A?

(d) Confirm that for the initial policy, given by $\boldsymbol{\theta}_0$, the value function is $v_{\pi_{\theta_0}}(A) = 10$ and $v_{\pi_{\theta_0}}(B) = 9$.

(e) If we in the update described in part (b) added a baseline equal to $v_{\pi_{\theta_0}}(s)$ what would the update be?

Remark: Of course, if we do not know the exact MDP we will not be able to use the unkown $v_{\pi_{\theta_0}}$ in the update, and would instead have to use some kind of function approximation.

3. In this problem we consider a one-step MDP (see slides from Lecture 8). There are two non-terminating states ("door A" and "door B"), and two actions ("open" and "don't open"). In each episode a state is drawn according to the state distribution $d(s)$, the agent chooses an action and receives a reward according to Table 1, and then the episode ends.

	open	don't open
door A	100	0
door B	0	100

Table 1: Rewards in Problem 3.

The agent uses a soft-max policy

$$\pi(a|s,\boldsymbol{\theta}) = \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,a))}{\sum_{b\in\mathcal{A}} \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}$$

with the features chosen as

$$\boldsymbol{x}(\text{door A}, \text{open}) = \boldsymbol{x}(\text{door B}, \text{open}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix},$$

$$\boldsymbol{x}(\text{door A}, \text{don't open}) = \boldsymbol{x}(\text{door B}, \text{don't open}) = \begin{bmatrix} 0 \\ 1 \end{bmatrix},$$

and $\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} \ln 10 \\ \ln 10 \end{bmatrix}$.

Note: Since $\boldsymbol{x}$ does not depend on s, the policy does not take into account which state we are in. This could be relevant if we have to choose the action without knowing if we stand in front of "door A" or "door B".

The initial state distribution $d(s)$ is given by $d(\text{door A}) = 0.8$ and $d(\text{door B}) = 0.2$, and we evaluate the agent based on the criterion

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_\theta}[v_{\pi_\theta}(s)] = \sum_s d(s) v_{\pi_\theta}(s) = \sum_{s,a,r} r p(r|s,a) \pi(a|s,\boldsymbol{\theta}) d(s).$$

We have seen in Lecture 8 that the gradient with respect to $\boldsymbol{\theta}$ is given by

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_\theta}[R \nabla \ln \pi(A|S,\boldsymbol{\theta})] = \sum_{s,a,r} [r \nabla \ln \pi(a|s,\boldsymbol{\theta})] p(r|s,a) \pi(a|s,\boldsymbol{\theta}) d(s).$$

Note: In all parts below, you can evaluate the expressions for $\boldsymbol{\theta} = \begin{bmatrix} \ln 10 \\ \ln 10 \end{bmatrix}$.

(a) Compute $\pi(a|s,\boldsymbol{\theta})$.

(b) Compute $\nabla \ln \pi(a|s,\boldsymbol{\theta})$.

(c) Compute the gradient $\nabla J(\boldsymbol{\theta})$. What will happen with the probabilities of choosing "open" or "don't open" if we update $\boldsymbol{\theta}$ using gradient ascent?

Hint: Note that there are four possible outcomes, see Table 1.

(d) Compute the gradient again, but this time use the expression with a baseline b,

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_\theta}[(R - b) \nabla \ln \pi(A|S,\boldsymbol{\theta})],$$

with $b = 50$.

(e) (*) In reinforcement learning we do not know $d(s)$ and $p(r|s,a)$ and can therefor not compute the full gradient. Instead we make an observation S, A, R and use an update on the form

$$\boldsymbol{\theta} \to \boldsymbol{\theta} + \alpha D,$$

where $D = \begin{bmatrix} d_1 \\ d_2 \end{bmatrix}$ is computed from S, A and R. From the two expressions we have seen for the full gradient, the following are potential candidates for D:

$$D_a = \begin{bmatrix} d_{a,1} \\ d_{a,2} \end{bmatrix} = R \nabla \ln \pi(A|S,\boldsymbol{\theta})$$

$$D_b = \begin{bmatrix} d_{b,1} \\ d_{b,2} \end{bmatrix} = (R - 50) \nabla \ln \pi(A|S,\boldsymbol{\theta})$$

As we have seen, $\mathbb{E}[D_a] = \mathbb{E}[D_b] = \nabla J(\boldsymbol{\theta})$, so both are unbiased estimates of the full gradient. However, the variances of D_a and D_b are not the same. Determine the variance of $d_{a,1}$ and $d_{b,1}$.

Hint: The formula for the variance is $\text{var}(d_1) = \mathbb{E}[(d_1 - \mathbb{E}[d_1])^2]$.

4. (*) Do Exercise 13.1 from the textbook. In this problem we will find the optimal policy for the ShortCorridor-example in Example 13.1 of the textbook. While the problem does not really does not use any theory on policy-gradients, it is a good exercise in using the Bellman equations to find the value function of a policy.

Hint: You can say that $\pi(left|s) = 1 - p$ and $\pi(right|s) = p$, where p is the probability of going right. Use the Bellman equations to set up a system of linear equations, and then solve this to find $v_\pi(0)$ (where state 0 is the starting state) as a function of p. Then find the p that maximize this value. At least try to determine the system of linear equations before you look at the solution!

Solutions

1.

(a) First note that

$$\begin{aligned}\ln \pi(a|s,\boldsymbol{\theta}) &= \ln \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,a)) - \ln \sum_{b\in\mathcal{A}} \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b)) \\ &= \boldsymbol{\theta}^\top \boldsymbol{x}(s,a) - \ln \sum_{b} \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b)).\end{aligned}$$

Taking the gradient of the first term gives

$$\nabla \left(\boldsymbol{\theta}^\top \boldsymbol{x}(s,a)\right) = \boldsymbol{x}(s,a).$$

For the second term we use

$$\nabla \ln f(\boldsymbol{\theta}) = \frac{\nabla f(\boldsymbol{\theta})}{f(\boldsymbol{\theta})},$$

so

$$\nabla \left(\ln \sum_{b} \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))\right) = \frac{\nabla \sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}.$$

Next we use

$$\nabla \exp(f(\boldsymbol{\theta})) = \exp(f(\boldsymbol{\theta}))\nabla f(\boldsymbol{\theta})$$

so we get

$$\frac{\nabla \sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))} = \frac{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))\boldsymbol{x}(s,b)}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))} = \sum_b \frac{\exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))\boldsymbol{x}(s,b)}{\sum_b \exp(\boldsymbol{\theta}^\top \boldsymbol{x}(s,b))} = \sum_b \pi(b|s,\boldsymbol{\theta})\boldsymbol{x}(s,b)$$

Hence we can conclude that

$$\nabla \ln \pi(a|s,\boldsymbol{\theta}) = \boldsymbol{x}(s,a) - \sum_{b\in\mathcal{A}} \pi(b|s,\boldsymbol{\theta})\boldsymbol{x}(s,b).$$

(b) We start by taking the logarithm

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

We first take the derivative with respect to $\boldsymbol{\theta}_\mu$ noting that $\nabla_{\theta_\mu}\mu(s,\boldsymbol{\theta}) = \boldsymbol{x}_\mu(s)$ and $\nabla_{\theta_\mu}\sigma(s,\boldsymbol{\theta}) = 0$ (since σ does not depend on $\boldsymbol{\theta}_\mu$).

$$\begin{aligned}\nabla_{\theta_\mu} \ln \pi(a|s,\boldsymbol{\theta}) &= -\frac{1}{2\sigma(s,\boldsymbol{\theta})^2}\nabla_{\theta_\mu}(a-\mu(s,\boldsymbol{\theta}))^2 \\ &= -\frac{1}{2\sigma(s,\boldsymbol{\theta})^2}\left[-2(a-\mu(s,\boldsymbol{\theta}))\nabla_{\theta_\mu}\mu(s,\boldsymbol{\theta})\right] \\ &= \frac{1}{\sigma(s,\boldsymbol{\theta})^2}(a-\mu(s,\boldsymbol{\theta}))\boldsymbol{x}_\mu(s).\end{aligned}$$

Next we take the derivative with respect to $\boldsymbol{\theta}_\sigma$ noting that $\nabla_{\theta_\sigma}\mu(s,\boldsymbol{\theta}) = 0$ and $\nabla_{\theta_\sigma}\sigma(s,\boldsymbol{\theta}) = \sigma(s,\boldsymbol{\theta})\boldsymbol{x}_\sigma(s,\boldsymbol{\theta})$,

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

Note that

$$\nabla_{\theta_\sigma}\frac{1}{\sigma(s,\boldsymbol{\theta})^2} = -\frac{2}{\sigma(s,\boldsymbol{\theta})^3}\nabla_{\theta_\sigma}\sigma(s,\boldsymbol{\theta}) = -\frac{2\sigma(s,\boldsymbol{\theta})\boldsymbol{x}_\sigma(s)}{\sigma(s,\boldsymbol{\theta})^3} = -\frac{2\boldsymbol{x}_\sigma(s)}{\sigma(s,\boldsymbol{\theta})^2},$$

and

$$\nabla_{\theta_\sigma} \ln \sigma(s, \boldsymbol{\theta}) = \frac{\nabla \sigma(s, \boldsymbol{\theta})}{\sigma(s, \boldsymbol{\theta})} = \boldsymbol{x}_\sigma(s).$$

Putting it together we get

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

2.

(a) Note that with the initial parameter vector we have $\boldsymbol{\theta}_0^\top \boldsymbol{x}(s, a) = 1$ for all s and a. Hence

$$\pi(a|s, \boldsymbol{\theta}_0) = \frac{e^1}{\sum_{b \in \mathcal{A}} e^1} = \frac{e^1}{2e^1} = \frac{1}{2}$$

since $\mathcal{A}$ contains two elements.

(b) We have already seen that the initial policy is $\pi(a|s, \boldsymbol{\theta}_0) = \frac{1}{2}$ for all s and a. The return after the first transition is

$$G_0 = 1 - 2 + 1 + 10 = 10.$$

Furthermore (using the solution to Problem 1a, or eq 13.9 in the textbook)

$$\nabla \ln \pi(A_0|S_0, \boldsymbol{\theta}_0) = \boldsymbol{x}(S_0, A_0) - \sum_b \pi(b|S_0, \boldsymbol{\theta}_0) \boldsymbol{x}(S_0, b) = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} - 0.5 \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} - 0.5 \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} -0.5 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix}.$$

The update is thus

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0 + \alpha G_0 \nabla \pi(A_0|S_0, \boldsymbol{\theta}_0) = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} + \underbrace{1}_{\alpha G} \begin{bmatrix} -0.5 \\ 0 \\ 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.5 \\ 1 \\ 1 \\ 1.5 \\ 1 \\ 1 \end{bmatrix}.$$

(c) With the new parameter vector we get

$$\boldsymbol{\theta}_1^\top \boldsymbol{x}(A, \texttt{Left}) = -0.5, \quad \boldsymbol{\theta}_1^\top \boldsymbol{x}(A, \texttt{Right}) = 1.5.$$

Hence,

$$\pi(\texttt{Right}|A, \boldsymbol{\theta}_1) = \frac{e^{1.5}}{e^{-0.5} + e^{1.5}} = 0.88$$

and thus $\pi(\texttt{Left}|A, \boldsymbol{\theta}_1) = 0.12$.

(d) The Bellman equations gives that

$$v_{\pi_{\theta_0}}(A) = 0.5(0 + v_{\pi_{\theta_0}}(A)) + 0.5(1 + v_{\pi_{\theta_0}}(B)) = 0.5 \times 10 + 0.5 \times 10 = 10.$$

and

$$v_{\pi_{\theta_0}}(B) = 0.5(-2 + v_{\pi_{\theta_0}}(A)) + 0.5(10 + v_{\pi_{\theta_0}}(C)) = 0.5 \times 8 + 0.5 \times 10 = 9.$$

Hence, the given function in the solves the Bellman equation and must thus be the true value function.

(e) Since in this case $G_0 - v_{\pi_{\theta_0}}(A) = 0$ we get

$$\boldsymbol{\theta}_1 = \boldsymbol{\theta}_0.$$

That is, since the return recieved was the same as the baseline we do not update anything.

3.

(a) We can note that for all actions and states we have $\boldsymbol{\theta}^\top \boldsymbol{x}(s,a) = e^{\ln 10} = 10$. Hence

$$\pi(a|s,\boldsymbol{\theta}) = \frac{10}{10+10} = \frac{1}{2},$$

for all a and s.

(b) From Problem 1 we have seen that

$$\nabla \ln \pi(a|s,\boldsymbol{\theta}) = \boldsymbol{x}(s,a) - \sum_{b\in\mathcal{A}} \pi(b|s,\boldsymbol{\theta})\boldsymbol{x}(s,b)$$

We note that

$$\sum_b \pi(b|s,\boldsymbol{\theta})\boldsymbol{x}(s,b) = \frac{1}{2}\begin{bmatrix}1\\0\end{bmatrix} + \frac{1}{2}\begin{bmatrix}0\\1\end{bmatrix} = \begin{bmatrix}1/2\\1/2\end{bmatrix}.$$

Hence, for both states

$$\nabla \ln \pi(\text{open}|s,\boldsymbol{\theta}) = \begin{bmatrix}1/2\\-1/2\end{bmatrix}, \quad \nabla \ln \pi(\text{don't open}|s,\boldsymbol{\theta}) = \begin{bmatrix}-1/2\\1/2\end{bmatrix}.$$

(c) We have four possible outcomes. Since the policy is uniformly random, the probability for ("door A", "open") and ("door A", "don't open") are both 0.4, while the probability for ("door B", "open") and ("door B", "don't open") are both 0.1. Hence

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_\theta}[R\nabla \ln \pi(A|S,\boldsymbol{\theta})] =$$
$$0.4\underbrace{\left(100\begin{bmatrix}1/2\\-1/2\end{bmatrix}\right)}_{\text{door A,open}} + 0.4\underbrace{\left(0\begin{bmatrix}-1/2\\1/2\end{bmatrix}\right)}_{\text{door A,don't open}} + 0.1\underbrace{\left(0\begin{bmatrix}1/2\\-1/2\end{bmatrix}\right)}_{\text{door B,open}} + 0.1\underbrace{\left(100\begin{bmatrix}-1/2\\1/2\end{bmatrix}\right)}_{\text{door B,don't open}} = \begin{bmatrix}15\\-15\end{bmatrix}$$

Hence, if we move along the positive gradient θ_1 will increase, while θ_2 will decrease. This in turn will give us a new policy that increases the probability of choosing "open" and decrease the probability of choosing "don't open". The reason for this is that the state "door A" has higher probability than "door B", and if the state is "door A" the best choice is "open".

(d) This time we get

$$\nabla J(\boldsymbol{\theta}) = \mathbb{E}_{\pi_\theta}[(R-b)\nabla \ln \pi(A|S,\boldsymbol{\theta})] =$$
$$0.4\underbrace{\left((100-50)\begin{bmatrix}1/2\\-1/2\end{bmatrix}\right)}_{\text{door A,open}} + 0.4\underbrace{\left((0-50)\begin{bmatrix}-1/2\\1/2\end{bmatrix}\right)}_{\text{door A,don't open}} +$$
$$0.1\underbrace{\left((0-50)\begin{bmatrix}1/2\\-1/2\end{bmatrix}\right)}_{\text{door B,open}} + 0.1\underbrace{\left((100-50)\begin{bmatrix}-1/2\\1/2\end{bmatrix}\right)}_{\text{door B,don't open}} = \begin{bmatrix}15\\-15\end{bmatrix}$$

which is the same as above.

(e) From part c and d we know that $\mathbb{E}[d_{a,1}] = \mathbb{E}[d_{b,1}] = 15$. Both $d_{a,1}$ and $d_{b,1}$ can take on the four possible values, as we have seen in part (c) and (d).

Let us start with $d_{a,1}$. It can take the following values:

$$\begin{aligned}(\text{door A, open}) &: 100/2 = 50\\ (\text{door A, don't open}) &: 0\\ (\text{door B, open}) &: 0\\ (\text{door B, don't open}) &: -100/2 = -50\end{aligned}$$

Hence the variance is

$$\text{var}(d_{a,1}) = 0.4(50-15)^2 + 0.4(0-15)^2 + 0.1(0-15)^2 + 0.1(-50-15)^2 = 1025.$$

With the baseline $b = 50$ we instead get

$$\begin{aligned}
\text{(door A, open)} &: (100-50)/2 = 25 \\
\text{(door A, don't open)} &: 50/2 = 25 \\
\text{(door B, open)} &: -50/2 = -25 \\
\text{(door B, don't open)} &: -(100-50)/2 = -25
\end{aligned}$$

Hence the variance is

$$\text{var}(d_{b,1}) = 0.4(25-15)^2 + 0.4(25-15)^2 + 0.1(-25-15)^2 + 0.1(-25-15)^2 = 400.$$

4. The Bellman equation for the value function states that (for $\gamma = 1$)

$$v_\pi(s) = \sum_a \pi(a|s) \sum_{s',r} p(s', r|s, a)[r + v_\pi(s')]$$

In our case, the environment is deterministic so $p(s', r|s, a)$ is either 0 or 1. Let us call the states $0, 1, 2, 3$ where 0 is the left state and 3 the terminating goal state. From the Bellman equation we thus get

$$v_\pi(0) = \underbrace{(1-p)}_{a=left}(-1 + v_\pi(0)) + \underbrace{p}_{a=right}(-1 + v_\pi(1)) = -1 + (1-p)v_\pi(0) + pv_\pi(1).$$

and in the same way for $v_\pi(1)$ and $v_\pi(2)$. We thus get the system of linear equations

$$\begin{aligned}
v_\pi(0) &= -1 + (1-p)v_\pi(0) + pv_\pi(1) \\
v_\pi(1) &= -1 + (1-p)v_\pi(2) + pv_\pi(0) \\
v_\pi(2) &= -1 + (1-p)v_\pi(1)
\end{aligned}$$

since $v_\pi(3) = 0$. We next insert $v_\pi(2)$ into the equation for $v_\pi(1)$ to get

$$v_\pi(1) = -1 + (1-p)(-1 + (1-p)v_\pi(1)) + pv_\pi(0) = -2 + p + (1-p)^2 v_\pi(1) + pv_\pi(0)$$

which we can rewrite as

$$(1-(1-p)^2)v_\pi(1) = -2 + p + pv_\pi(0) \iff v_\pi(1) = \frac{-2 + p + pv_\pi(0)}{2p - p^2}$$

Finally, the equation for $v_\pi(0)$ can be rewritten as

$$pv_\pi(0) = -1 + pv_\pi(1) = -1 + \frac{-2 + p + pv_\pi(0)}{2-p} = -1 + \frac{-2+p}{2-p} + \frac{p}{2-p}v_\pi(0) = -2 + \frac{p}{2-p}v_\pi(0)$$

which we can rewrite as

$$\left(p - \frac{p}{2-p}\right)v_\pi(0) = -2 \iff \frac{p(1-p)}{2-p}v_\pi(0) = -2 \iff v_\pi(0) = -\frac{2(2-p)}{p(1-p)}$$

We can note that if we let $p \to 0$ or $p \to 1$ the value goes to $-\infty$, since a deterministic policy that takes the same action in all states will not be able to reach the goal.

One way of finding the p that maximize this is to set the derivative w.r.t p to 0. You will see that this gives you $p = 2 \pm \frac{\sqrt{8}}{2}$. Since p should be a probability, the solution must be $p = 2 - \frac{\sqrt{8}}{2} = 2 - \sqrt{2} \approx 0.5858$. Using this p we thus get

$$v_\pi(0) \approx -11.65.$$