PROBABILISTIC ROBOTICS:
PARTIALLY OBERVABLE MARKOV DECISION PROCESSES

Pierre-Paul TACHER

1.

1.1. We define the states

x_1**:** the tiger is behind door 1.
x_2**:** the tiger is behind door 2.

The associated belief state space is

$$\begin{aligned} b &= (p_1, p_2) \\ &= (p_1, 1 - p_1) \end{aligned}$$

where p_1 is the probability that tiger is behind door 1. We define the actions:

u_1**:** open door 1.
u_2**:** open door 2.
u_3**:** listen.

The rewards incurred by actions :

$$\begin{aligned} r(b, u_1) &= \mathrm{E}_x[r(x, u_1)] \\ &= p(x = x_1) \times (-20) + p(x = x_2) \times 10 \\ &= -20p_1 + 10(1 - p_1) \\ r(b, u_2) &= \mathrm{E}_x[r(x, u_2)] \\ &= p(x = x_1) \times 10 + p(x = x_2) \times (-20) \\ &= 10p_1 - 20(1 - p_1) \\ \forall b, \quad r(b, u_3) &= -1 \end{aligned}$$

Let $Z \in \{z_1, z_2\}$ the random variable which modelizes the measurement.

z_1**:** the roar seems to come from behind door 1.
z_2**:** the roar seems to come from behind door 2.

i	$P(Z = z_i \mid x_1)$
1	0.85
2	0.15

i	$P(Z = z_i \mid x_2)$
1	0.15
2	0.85

1.2. Let $\pi = (u_3, u_3, u_1)$ a tree steps open loop policy. The reward computed in the belief state obtained by executing this policy is

$$R_3(b_0, \pi) = r(b_0, u_3) + r(b_1, u_3) + r(b_2, u_1)$$

where $b_0 = (p_1, 1-p_1)$, b_1 and b_2 are belief states obtained after each listening action; They themselves are random variables. We take the expectancy over them to get

$$\begin{aligned} V_3(b_0, \pi) &= \mathrm{E}_{b_1,b_2}[R_3(b_0, \pi)] \\ &= r(b_0, u_3) + \mathrm{E}_{b_1,b_2}[r(b_1, u_3)] + \mathrm{E}_{b_1,b_2}[r(b_2, u_1)] \\ &= -1 + \mathrm{E}_{b_1,b_2}[-1] + \mathrm{E}_{b_1,b_2}[r(b_2, u_1)] \\ &= -2 + \mathrm{E}_{b_1,b_2}[r(b_2, u_1)] \qquad \textcircled{1} \end{aligned}$$

We note $B(b_0, \pi, z_i, z_j), \quad (i,j) \in \{1,2\}^2$ the distribution / belief state aquired from b_0 after executing the first two steps of policy π, after sensing z_i and then z_j. Under this distribution, the probability that the tiger is behind door 1 is

$$\begin{aligned} B(b_0, \pi, z_i, z_j)(x = x_1) &= p(x = x_1 \mid (Z_1, Z_2) = (z_i, z_j)) \\ &= \frac{p(x = x_1 \cap (Z_1, Z_2) = (z_i, z_j))}{p((Z_1, Z_2) = (z_i, z_j))} \end{aligned}$$

with

$$\begin{aligned} p(x = x_1 \cap (Z_1, Z_2) = (z_i, z_j)) &= p(Z_2 = z_j \mid x = x_1, Z_1 = z_i) \times p(x = x_1 \mid Z_1 = z_i) \\ &= p(Z_2 = z_j \mid x = x_1) \times p(Z_1 = z_i \mid x = x_1) \times p(x = x_1) \\ &= p(Z_2 = z_j \mid x = x_1) \times \frac{p(Z_1 = z_i \mid x = x_1) \times p_1}{p(Z_1 = z_i)} \\ &= p(Z_2 = z_j \mid x = x_1) \times \frac{p(Z_1 = z_i \mid x = x_1) \times p_1}{p(Z_1 = z_i \cap x_1) + p(Z_1 = z_i \cap x_2)} \\ &= p(Z_2 = z_j \mid x = x_1) \times \frac{p(Z_1 = z_i \mid x = x_1) \times p_1}{p(Z_1 = z_i \mid x_1)p_1 + p(Z_1 = z_i \mid x_2)(1-p_1)} \end{aligned}$$

$$\begin{aligned} &p((Z_1, Z_2) = (z_i, z_j)) \\ &= p(Z_2 = z_j \mid Z_1 = z_i) \times p(Z_1 = z_i) \\ &= p(Z_2 = z_j \mid Z_1 = z_i) \times p(Z_1 = z_i) \\ &= \big[p(Z_2 = z_j \cap x_1 \mid Z_1 = z_i) + p(Z_2 = z_j \cap x_2 \mid Z_1 = z_i)\big] \times p(Z_1 = z_i) \\ &= \big[p(Z_2 = z_j \mid x_1, Z_1 = z_i) \times p(x_1 \mid Z_1 = z_i) + p(Z_2 = z_j \mid x_2, Z_1 = z_i) \times p(x_2 \mid Z_1 = z_i)\big] \\ &\quad \times p(Z_1 = z_i) \\ &= \big[p(Z_2 = z_j \mid x_1) \times \frac{p(Z_1 = z_i \mid x_1)p_1}{p(Z_1 = z_i)} + p(Z_2 = z_j \mid x_2) \times \frac{p(Z_1 = z_i \mid x_2)(1-p_1)}{p(Z_1 = z_i)}\big] \\ &\quad \times p(Z_1 = z_i) \\ &= p(Z_2 = z_j \mid x_1) \times p(Z_1 = z_i \mid x_1)p_1 + p(Z_2 = z_j \mid x_2) \times p(Z_1 = z_i \mid x_2)(1-p_1) \end{aligned}$$

Actually we don't need the last 2 calculations to compute ①; the probability that the distribution b_2 would be $B(b_0, \pi, z_i, z_j)$ is $p((Z_1, Z_2) = (z_i, z_j))$, so:

$$\begin{aligned} \mathrm{E}_{b_1,b_2}[r(b_2, u_1)] &= \sum_{(i,j)\in\{1,2\}^2} p((Z_1, Z_2) = (z_i, z_j)) \times \big[-20B(b_0, \pi, z_i, z_j)(x_1) + 10(1 - B(b_0, \pi, z_i, z_j)(x_1))\big] \\ &= \sum_{(i,j)\in\{1,2\}^2} -20p(x = x_1 \cap (Z_1, Z_2) = (z_i, z_j)) + 10\big[p((Z_1, Z_2) = (z_i, z_j)) - p(x = x_1 \cap (Z_1, Z_2) = (z_i, z_j))\big] \\ &= -20p(x = x_1) + 10(1 - p(x = x_1)) \\ &= V_1(b_0, u_1) \end{aligned}$$

the last value being the 1 step horizon expected gain of choosing action u_1.

$$V_3(b_0, \pi) = -2 + V_1(b_0, u_1)$$

These somewhat contrived calculations show that we wasted a cost of 2 for listening before eventually opening door 1 regardless of the information gained by sensing; we had better just open door 1 in the first place.

1.3. Now the policy is $\Pi = (u_3, f(Z_1))$ with
The reward computed in the belief state obtained by executing this policy is

z	$f(z)$
z_1	u_2
z_2	u_1

$$\begin{aligned} R_2(b_0, \Pi) &= r(b_0, u_3) + r(b_1, f(Z_1)) \\ &= -1 + r(b_1, f(Z_1)) \end{aligned}$$

We take the expectancy over distribution b_1 to get

$$\begin{aligned} V_2(b_0, \Pi) &= -1 + \mathrm{E}_{b_1}[r(b_1, f(Z_1))] \\ &= -1 + \sum_{i=1}^{2} p(Z_1 = z_i) \times r(B(b_0, z_i), f(z_i)) \\ &= -1 + p(Z_1 = z_1) \times r(B(b_0, z_1), u_2) + p(Z_1 = z_2) \times r(B(b_0, z_2), u_1) \end{aligned}$$

We have :

$$\begin{aligned} B(b_0, z_1)(x_1) &= p(x_1 \mid z_1) \\ &= \frac{p(z_1 \mid x_1)}{p(z_1)} \times p_1 \\ &= \frac{0.85}{0.85p_1 + 0.15(1 - p_1)} \times p_1 \end{aligned}$$

$$\begin{aligned} B(b_0, z_2)(x_1) &= p(x_1 \mid z_2) \\ &= \frac{p(z_2 \mid x_1)}{p(z_2)} \times p_1 \\ &= \frac{0.15}{0.15p_1 + 0.85(1 - p_1)} \times p_1 \end{aligned}$$

$$\begin{aligned} V_2(b_0, \Pi) &= -1 + \big[10 \times 0.85 \times p1 - 20 \times 0.15 \times (1 - p_1)\big] + \big[- 20 \times 0.15 \times p1 + 10 \times 0.85 \times (1 - p_1)\big] \\ &= -1 + 10 \times 0.85 - 20 \times 0.15 \\ &= 4.5 \end{aligned}$$

This does not depend on the initial belief b_0.

1.4. We recall

$$\begin{aligned} r(b, u_1) &= -20p_1 + 10(1 - p_1) \\ r(b, u_2) &= 10p_1 - 20(1 - p_1) \\ \forall b, \quad r(b, u_3) &= -1 \end{aligned}$$

They are represented in figure 1 along with the optimal expected payoff

$$V_1(b_0) = \max_u r(b_0, u)$$

$$V_1(b_0) = \max \left\{ \begin{array}{c} -20p_1 + 10(1 - p_1) \\ 10p_1 - 20(1 - p_1) \\ -1 \end{array} \right\} \begin{array}{l} (*) \\ (*) \\ (*) \end{array}$$

Only the linear equations marked with $(*)$ contribute.

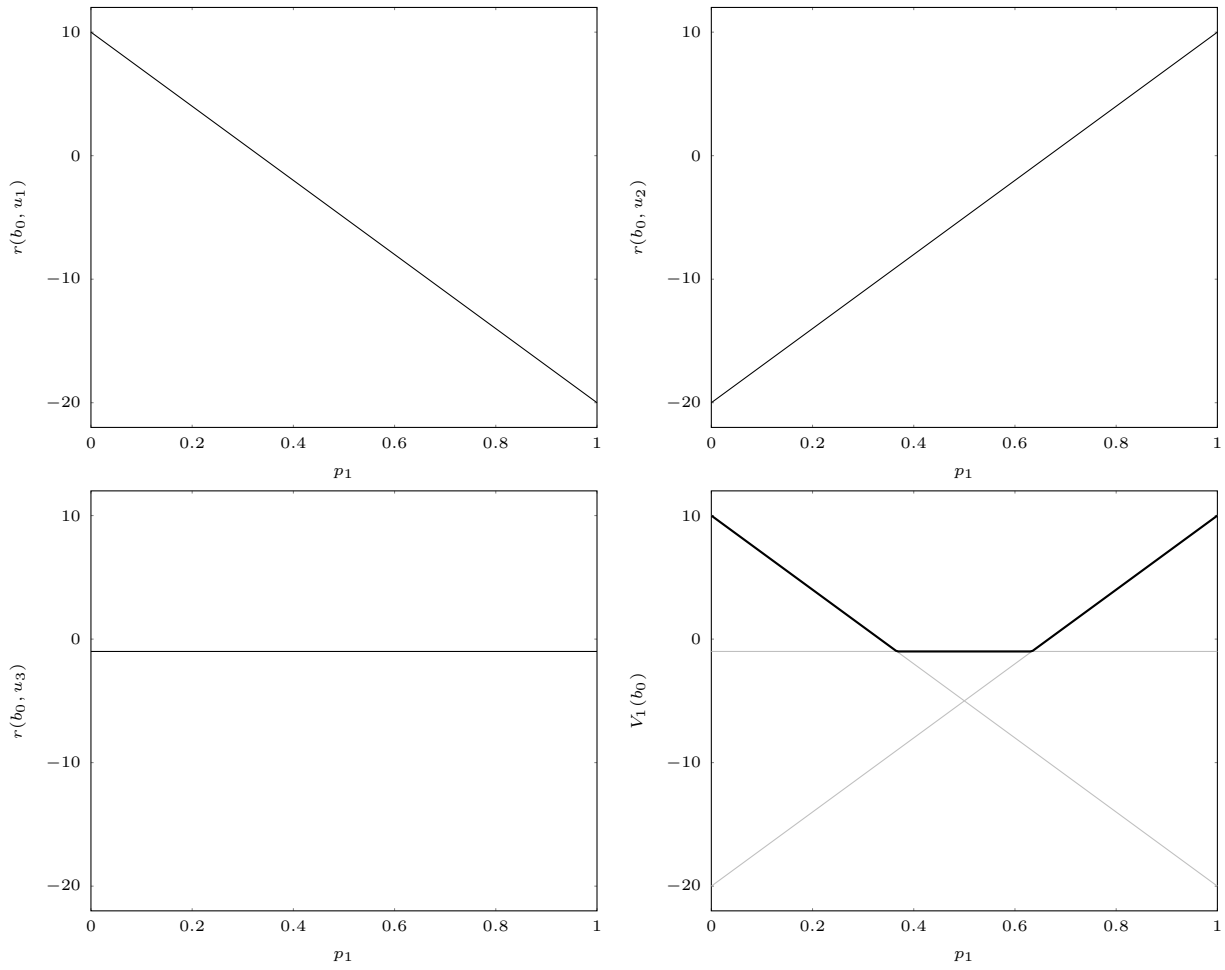

FIGURE 1. 1 step horizon expected rewards

1.5. Suppose we begin to listen (u_3) and sense z_1. The belief evolves according to:

$$\begin{aligned} B(b_0, z_1)(x = x_1) &= p(x_1 \mid z_1) \\ &= \frac{p(z_1 \mid x_1)p_1}{p(z_1)} \\ &= \frac{0.85p_1}{0.85p_1 + 0.15(1 - p_1)} \end{aligned}$$

The 2 steps horizon expected reward will then be

$$\begin{aligned} V_1(B(b_0, z_1)) &= \max \left\{ \begin{array}{c} -20\dfrac{0.85p_1}{0.85p_1 + 0.15(1 - p_1)} + 10\dfrac{0.15(1 - p_1)}{0.85p_1 + 0.15(1 - p_1)} \\ 10\dfrac{0.85p_1}{0.85p_1 + 0.15(1 - p_1)} - 20\dfrac{0.15(1 - p_1)}{0.85p_1 + 0.15(1 - p_1)} \\ -1 \end{array} \right\} \\ &= \frac{1}{0.85p_1 + 0.15(1 - p_1)} \max \left\{ \begin{array}{c} -20 \times 0.85p_1 + 10 \times 0.15(1 - p_1) \\ 10 \times 0.85p_1 - 20 \times 0.15(1 - p_1) \\ -0.85p_1 - 0.15(1 - p_1) \end{array} \right\} \end{aligned}$$

See figure 2 for related graphs. Similarly, suppose we sense z_2. The belief evolves according to:

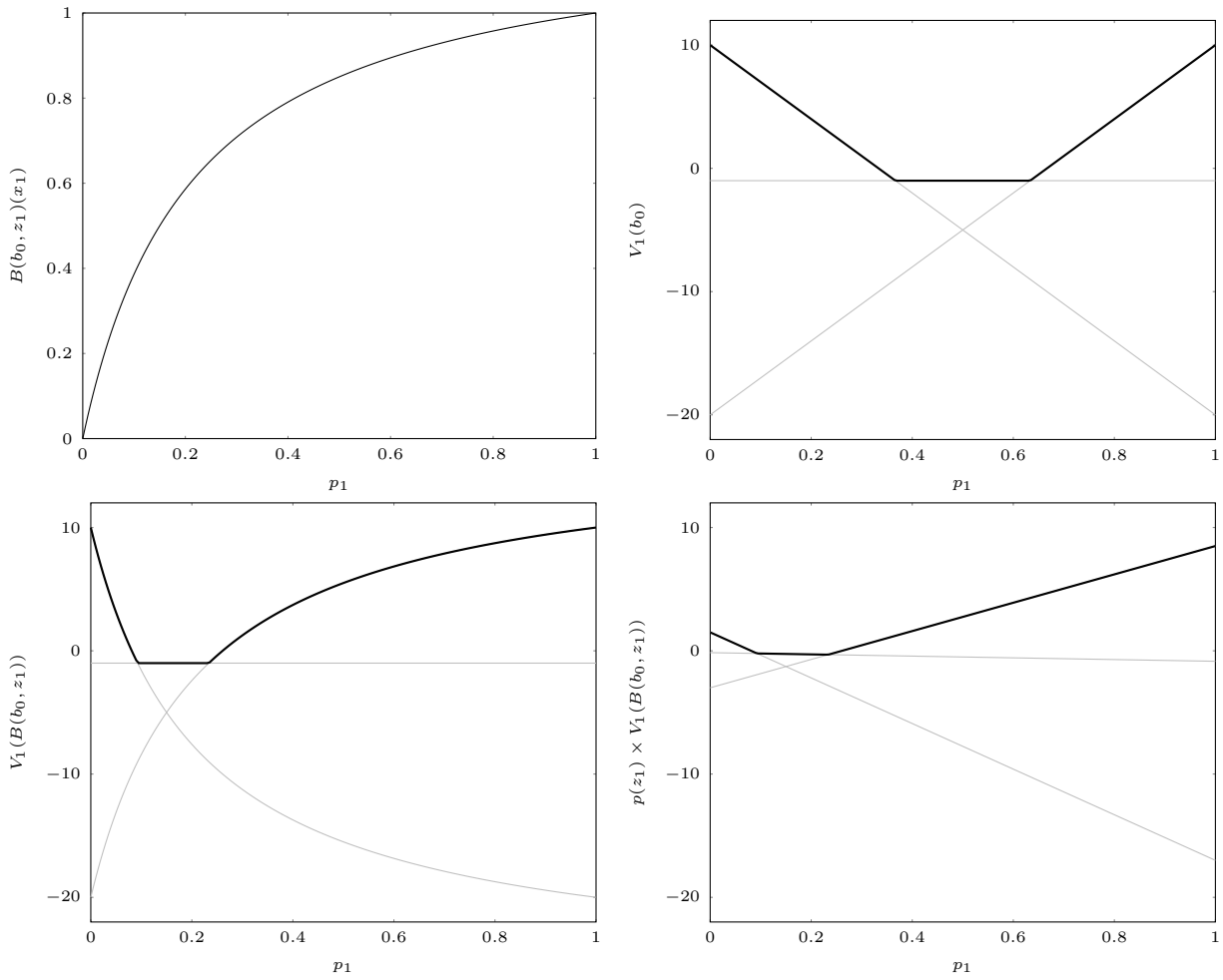

FIGURE 2. expected rewards after having sensed z_1

$$\begin{aligned} B(b_0, z_2)(x = x_1) &= p(x_1 \mid z_2) \\ &= \frac{p(z_2 \mid x_1)p_1}{p(z_2)} \\ &= \frac{0.15p_1}{0.15p_1 + 0.85(1 - p_1)} \end{aligned}$$

The 2 steps horizon expected reward will then be

$$\begin{aligned} V_1(B(b_0, z_2)) &= \max \left\{ \begin{array}{c} -20\dfrac{0.15p_1}{0.15p_1 + 0.85(1 - p_1)} + 10\dfrac{0.85(1 - p_1)}{0.15p_1 + 0.85(1 - p_1)} \\ 10\dfrac{0.15p_1}{0.15p_1 + 0.85(1 - p_1)} - 20\dfrac{0.85(1 - p_1)}{0.15p_1 + 0.85(1 - p_1)} \\ -1 \end{array} \right\} \\ &= \frac{1}{0.15p_1 + 0.85(1 - p_1)} \max \left\{ \begin{array}{c} -20 \times 0.15p_1 + 10 \times 0.85(1 - p_1) \\ 10 \times 0.15p_1 - 20 \times 0.85(1 - p_1) \\ -0.15p_1 - 0.85(1 - p_1) \end{array} \right\} \end{aligned}$$

See figure 3 for related graphs. We now integrate the cost of u_3, and compute the expectancy over Z_1 to

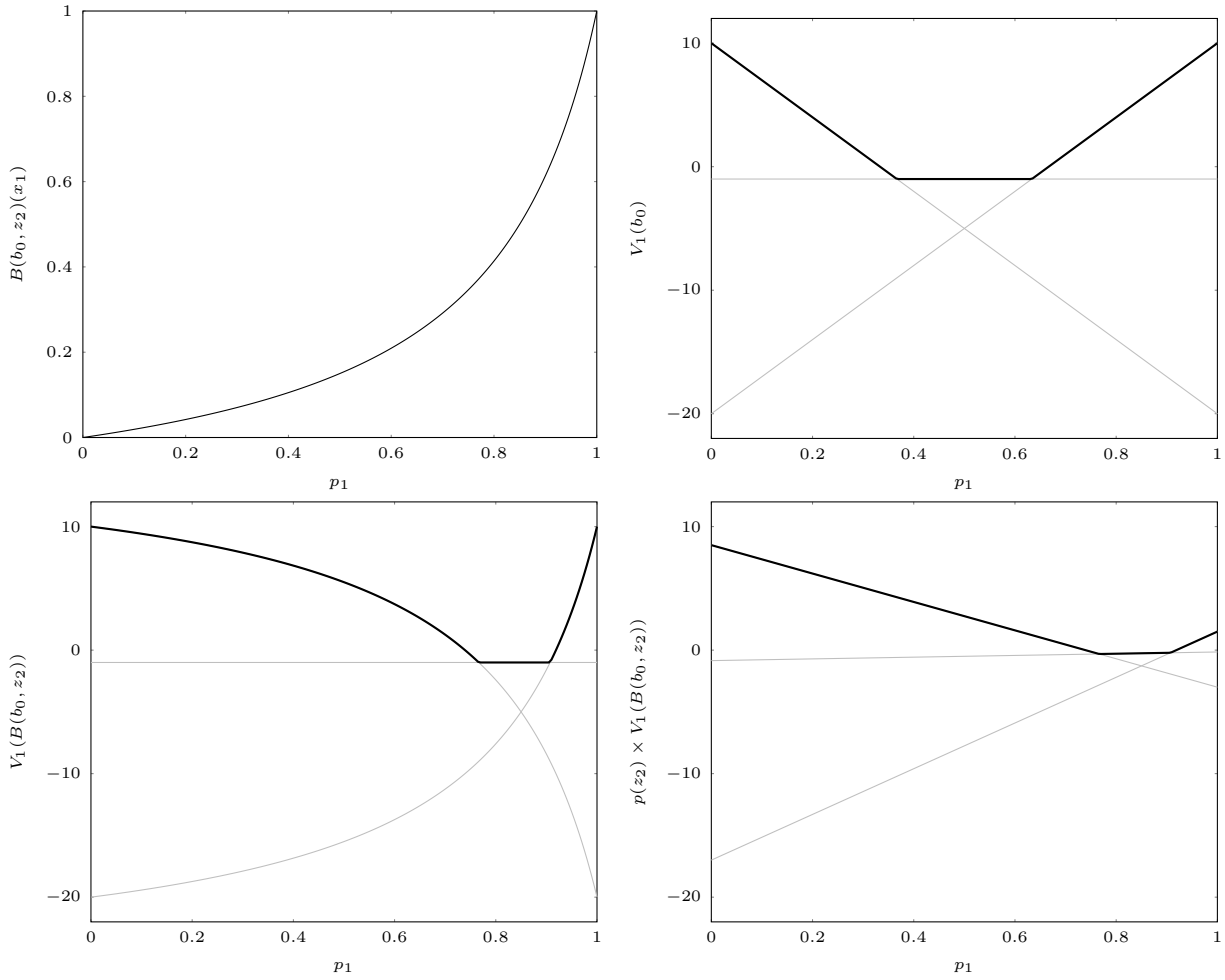

FIGURE 3. expected rewards after having sensed z_2

have the full 2 steps horizon expected reward when choosing to listen first:

$$
\begin{aligned}
&V_2(b_0, u_3) \\
&= -1 + p(Z_1 = z_1) \times V_1(B(b_0, z_1)) + p(Z_1 = z_2) \times V_1(B(b_0, z_2)) \\
&= -1 + \max \left\{ \begin{array}{c} -20 \times 0.85 p_1 + 10 \times 0.15(1 - p_1) \\ 10 \times 0.85 p_1 - 20 \times 0.15(1 - p_1) \\ -0.85 p_1 - 0.15(1 - p_1) \end{array} \right\} + \max \left\{ \begin{array}{c} -20 \times 0.15 p_1 + 10 \times 0.85(1 - p_1) \\ 10 \times 0.15 p_1 - 20 \times 0.85(1 - p_1) \\ -0.15 p_1 - 0.85(1 - p_1) \end{array} \right\} \\
&= -1 + \max \left\{ \begin{array}{c} -30 p_1 + 10 \\ -15.5 \\ -17.15 p_1 + 0.65(1 - p_1) \\ 5.5 \\ 30 p_1 - 20 \\ 8.35 p_1 - 3.85(1 - p_1) \\ -3.85 p_1 + 8.35(1 - p_1) \\ 0.65 p_1 - 17.15(1 - p_1) \\ -1 \end{array} \right\}
\end{aligned}
$$

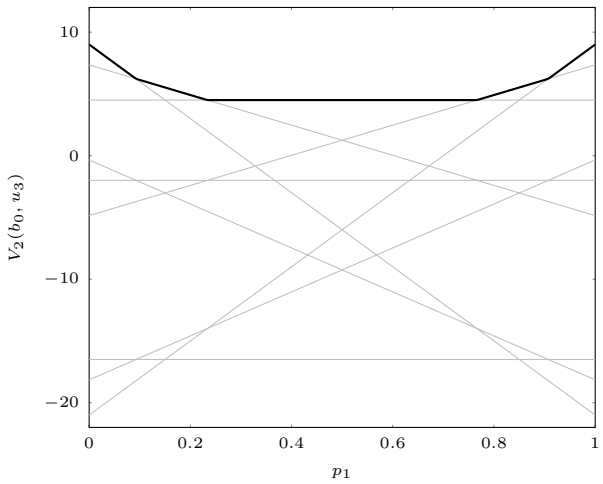

FIGURE 4. optimal expected rewards when choosing u_3 as first action

$$
\begin{aligned}
&= \max \left\{ \begin{array}{c} -30p_1 + 9 \\ -16.5 \\ -18.15p_1 - 0.35(1-p_1) \\ 4.5 \\ 30p_1 - 21 \\ 7.35p_1 - 4.85(1-p_1) \\ -4.85p_1 + 7.35(1-p_1) \\ -0.35p_1 - 18.15(1-p_1) \\ -2 \end{array} \right\} \begin{array}{l} (*) \\ \\ \\ (*) \\ (*) \\ (*) \\ (*) \\ \\ \end{array} \\
&= \max \left\{ \begin{array}{c} -30p_1 + 9 \\ 4.5 \\ 30p_1 - 21 \\ 7.35p_1 - 4.85(1-p_1) \\ -4.85p_1 + 7.35(1-p_1) \end{array} \right\}
\end{aligned}
$$

$V_2(b_0, u_3)$ is represented figure 4. Now we have to integrate the possible choices u_1 and u_2 as first action:

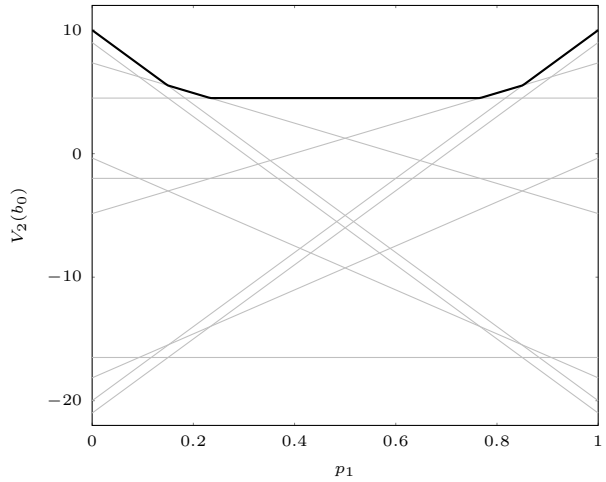

FIGURE 5. horizon 2 optimal expected rewards

$$
\begin{aligned}
&V_2(b_0)\\
&= \max \left\{ \begin{array}{cc}
-20p_1 + 10(1-p_1) & (*)\\
10p_1 - 20(1-p_1) & (*)\\
-30p_1 + 9 & \\
-16.5 & \\
-18.15p_1 - 0.35(1-p_1) & \\
4.5 & (*)\\
30p_1 - 21 & \\
7.35p_1 - 4.85(1-p_1) & (*)\\
-4.85p_1 + 7.35(1-p_1) & (*)\\
-0.35p_1 - 18.15(1-p_1) & \\
-2 &
\end{array} \right\}\\
&= \max \left\{ \begin{array}{c}
-20p_1 + 10(1-p_1)\\
10p_1 - 20(1-p_1)\\
4.5\\
7.35p_1 - 4.85(1-p_1)\\
-4.85p_1 + 7.35(1-p_1)
\end{array} \right\}
\end{aligned}
$$

$V_2(b_0)$ is represented figure 5