Problem 1:

$V^{\pi_D}(s) = Q^{\pi_D}(s, \pi_D(s)). \quad \forall s \in N.$

$Q^{\pi_D}(s, \pi_D(s)) = R(s, \pi_D(s)) + \sum_{s' \in N} P(s, \pi_D(s), s') V^{\pi_D}(s'). \quad \forall s \in N.$

$V^{\pi_D}(s) = R^{\pi_D}(s) + \sum_{s' \in N} P^{\pi_D}(s, s') V^{\pi_D}(s'). \quad \forall s \in N.$

$Q^{\pi_D}(s, \pi_D(s)) = R(s, \pi_D(s)) + \sum_{s' \in N} P(s, \pi_D(s), s') Q^{\pi_D}(s', \pi_D(s')). \quad \forall s \in N.$

Problem 2:

1. Leave the current state : $P = a$, $R = 1-a$.

2. Stay at the current state : $P = 1-a$, $R = 1+a$.

Therefore, the reward and the transition probability depends solely on a. Let state be : {leave, stay}. Then, whether the current state is leave or stay, both of them has exactly the same probability transitions and reward functions. Therefore, they will have the same value functions. That is:

P	leave	stay
leave	a	1-a
stay	a	1-a.

R	leave	stay.
leave	1-a	1+a
stay	1-a	1+a.

Let S_1 = leave. S_2 = stay.

$V(S_1) = a(1-a) + (1-a)(1+a) + r a V(S_1) + r(1-a) V(S_2)$, for any a.

$V(S_2) = a(1-a) + (1-a)(1+a) + r a V(S_1) + r(1-a) V(S_2).$

Thus : $V^*(S_1) = V^*(S_2)$, and:

$$V^*(S_1) = \max_{a \in [0,1]} a(1-a) + (1-a)(1+a) + r a V^*(S_1) + r(1-a) V^*(S_1).$$

$$= \max_{a \in [0,1]} a(1-a) + (1-a)(1+a) + r V^*(S_1)$$

Required: $1 - 2a - 2a = 0 \Rightarrow a = \frac{1}{4}$. Thus optimal policy for all state : $a = \frac{1}{4}$.

$0.5 V^*(S_1) = \frac{3}{4} \times \frac{3}{2} = \frac{9}{8} \Rightarrow V^*(S_1) = V^*(S_2) = \frac{9}{4}.$

Therefore, under optimal policy $\alpha = \frac{1}{4}$, at current state s, the choice of leaving and staying gives the same result. Therefore, $V^*(s) = \frac{9}{4}$ $\forall s \in S$. $\pi^*(s) = \frac{1}{4}$ $\forall s \in S$.

Problem 3:

State space : $S = \{0, \cdots, n\}$, $N = \{1, \cdots, n-1\}$, $T = \{0, n\}$. Representing lilypads.

Action Space : $A = \{A, B\}$. Representing sounds.

$$P_R(s, a, r, s') = \begin{cases} \frac{s}{n}, & \text{if } a = A,\ r = -1,\ s' = s-1. \\ \frac{n-s}{n}, & \text{if } a = A,\ r = 1,\ s' = s+1 \\ \frac{1}{n}, & \text{if } a = B,\ r = s'-s,\ s' \in S. \\ 0, & \text{otherwise} \end{cases}$$

Transitions function:

$P(s, A, s+1) = \frac{n-s}{n}$, $P(s, A, s-1) = \frac{s}{n}$, $P(s, B, i) = \frac{1}{n}$, $\forall i \in S$, $\forall s \in N$.

Reward function :

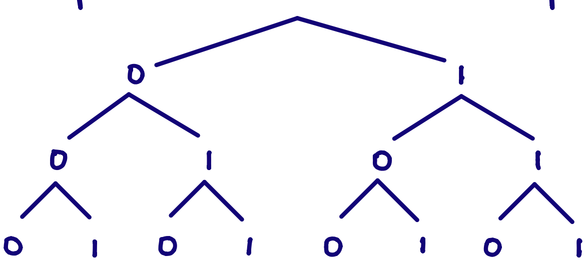
$R(s, A) = 1 \times \frac{n-s}{n} - \frac{s}{n} = \frac{n-2s}{n}$. $\forall s \in N$.

$R(s, B) = \frac{1}{n}\left(\sum_{i=0}^{n} i\right) = \frac{1}{n} \cdot (n)(n+1)\frac{1}{2} = \frac{n+1}{2}$. $\forall s \in N$.

How we get 2^n possible deterministic policy : For each s, we can choose A or B.

Algorithm used to get all possible deterministic policy :

Example: Binomial Tree $\Rightarrow$ Express in list :



							leaves							
index	0	1	2	3	4	5	6	7	8	9	10	11	12	13
value [	0	1	0	1	0	1	0	1	0	1	0	1	0	1]

node.index = floor (leaf. index / 2) - 1.

If index = even, value = 0. If index = odd, value = 1. Therefore :

For all leaves (leaves ends at : $\sum_{i=1}^{n} 2^i$, leaves starts at : $\sum_{i=1}^{n-1} 2^i$) :

1) If leaf. index = even, add 0, else add 1.

2) If floor (leaf. index / 2) - 1 = even add 0, end add 1.

3) Set leaf = floor(leaf.index/2) - 1, repeat 2) n-1 times.

Although the order is reverted, by symmetry, reverting it is not necessary.

Problem 4:

$P(s,a,s') = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s'-s)^2}{2\sigma^2}}$, $R(s,a,s') = -e^{as'}$

$$V^*(s) = \max_{a\in\mathbb{R}} R(s,a)$$
$$= \max_{a\in\mathbb{R}} \int_{s'\in N} P(s,a,s') R(s,a,s')\, ds'$$
$$= \max_{a\in\mathbb{R}} \int_{s'\in N} -e^{as'} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s'-s)^2}{2\sigma^2}} ds'.$$
$$= \max_{a\in\mathbb{R}} \int_{s'\in N} -\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s'-s)^2}{2\sigma^2}+as'} ds'.$$
$$= \max_{a\in\mathbb{R}} \int_{s'\in N} -\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{s'^2-2s's+s^2-2\sigma^2 as'}{2\sigma^2}} ds'$$
$$= \max_{a\in\mathbb{R}} \int_{s'\in N} -\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{s'^2-2(s+\sigma^2 a)s'+(s+\sigma^2 a)^2-2\sigma^2 as-\sigma^4 a^2}{2\sigma^2}} ds'$$
$$= \max_{a\in\mathbb{R}} -e^{as+\frac{\sigma^2 a^2}{2}} \int_{s'\in N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(s'-(s+\sigma^2 a))^2}{2\sigma^2}} ds'.$$
$$= \max_{a\in\mathbb{R}} -e^{as+\frac{\sigma^2 a^2}{2}}.$$

Thus, we requires : $-e^{as+\frac{\sigma^2 a^2}{2}}(s+\sigma^2 a) = 0. \Rightarrow s+\sigma^2 a = 0. \Rightarrow a = -\frac{s}{\sigma^2}.$

Thus, optimal action for state s : $a = -\frac{s}{\sigma^2}$. Optimal cost is: $-e^{-\frac{s^2}{\sigma^2}+\frac{s^2}{2\sigma^2}} = -e^{-\frac{s^2}{2\sigma^2}}$.