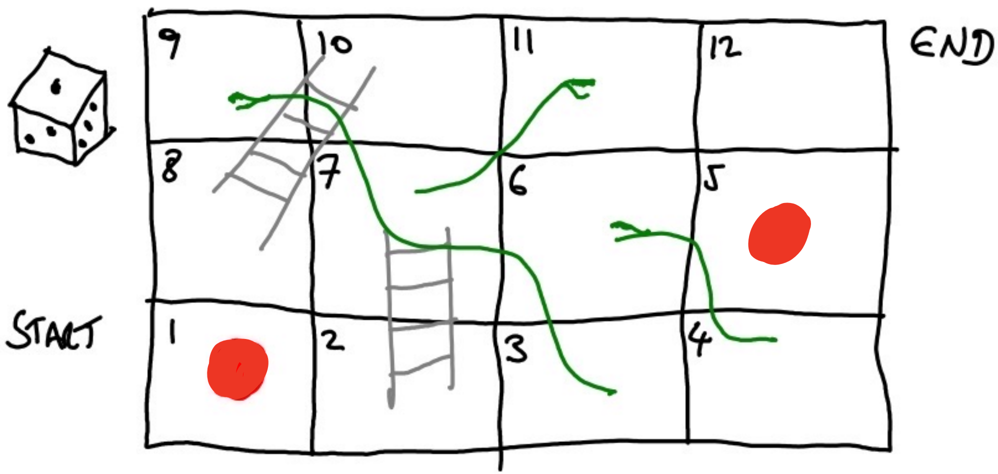
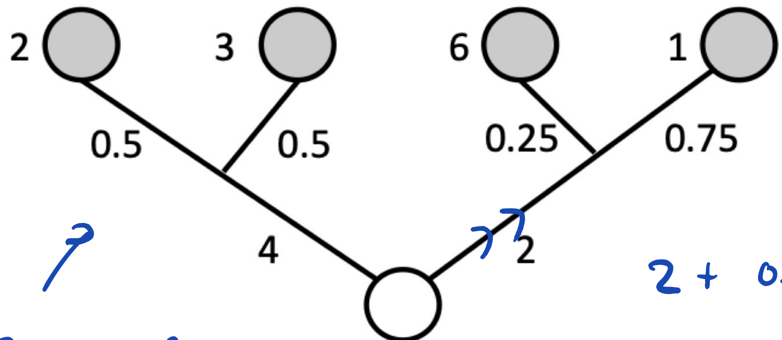
3. MARKOV DECISION PROCESSES

A MARKOV CHAIN WITH DECISIONS:



[CHOOSE WHICH TO MOVE]

SOLVING AN MDP:



$4 + 2 \times 0.5 + 3 \times 0.5$

$= 4 + 2.5$

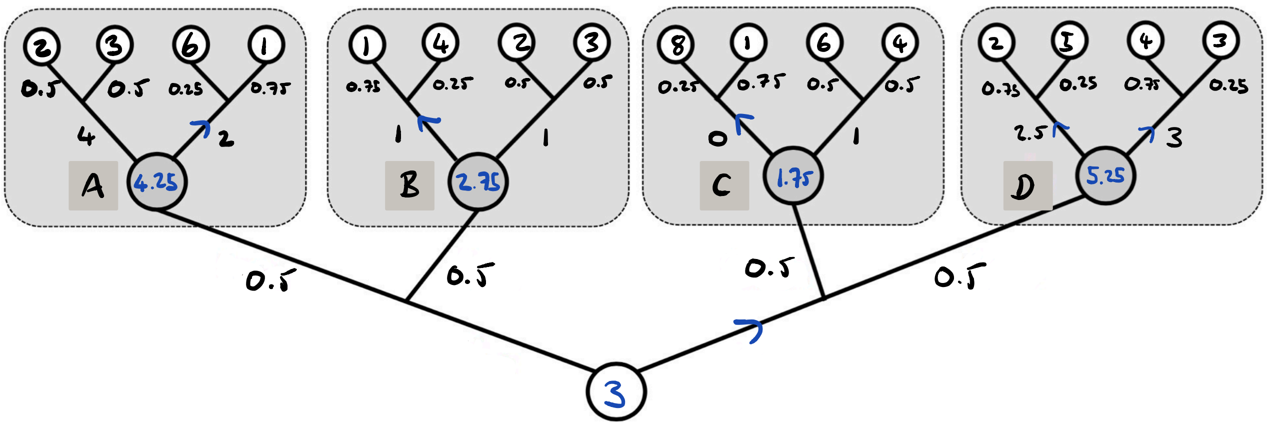
$= 6$

$2 + 0.25 \times 6 + 0.75 \times 1$

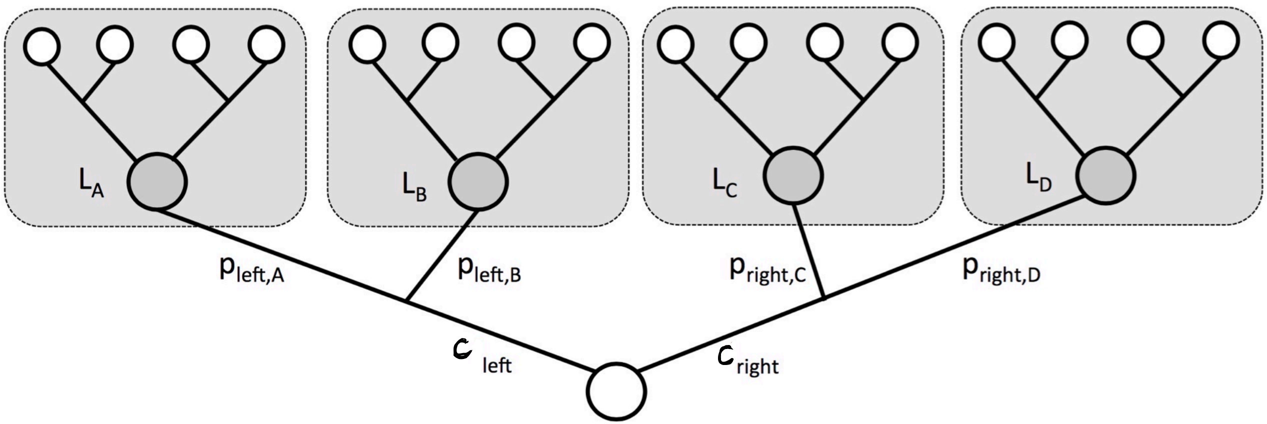
$= 2 + 2.25$

$= 4.25$

SOLVING AN MDP:

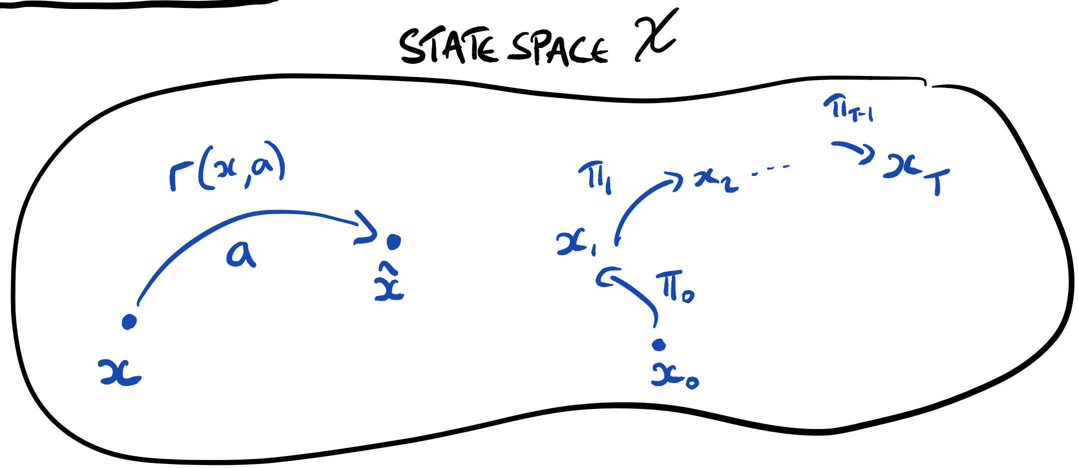


SOLVING AN MDP:



$$L(x) = \min_{a \in \{\text{LEFT}, \text{RIGHT}\}} \left\{ c(a) + \mathbb{E}_{x,a}\left[L(\hat{x}) \right] \right\}$$

ABSTRACT DEFINITION:



STATE $\quad x \in \mathcal{X}$

ACTION $\quad a \in \mathcal{A}$

REWARD $\quad r(x,a)$

NEXT STATE

$$\hat{x} = f(x, a, u)$$

POLICY $\quad \pi(x)$ OR $\pi_t(x)$.

OBJECTIVE: [MAXIMIZE SUM OF REWARDS]

$$V_T(x_0) := \max \underbrace{\mathbb{E}_{x_0}\left[\sum_{t=0}^{T-1} r(x_t, \pi_t) + r(x_T)\right]}_{R_T(x,\pi)} \text{ over } \pi \in \mathcal{P}_t$$

VALUE FUNCTION

[POLICY OVER t STEPS]

THE BELLMAN EQUATION

THEOREM:
$$V_t(x) = \max_{a \in A} \left\{ r(x,a) + \mathbb{E}_{x,a}\left[V_{t-1}(\hat{x})\right]\right\}, \qquad V_0(x) = r(x)$$

HERE
$$\mathbb{E}_{x,a}\left[V(\hat{x})\right] = \mathbb{E}\left[V(X_1) \mid X_0 = x, A_0 = a\right] = \mathbb{E}\left[V(f(x,a,u))\right]$$

LEMMA: a) $\mathcal{P}_t(x) = \mathcal{A} \times \mathcal{P}_{t-1}$
$$\therefore \max_{\pi \in \mathcal{P}_t(x)} f(\pi) = \max_{a} \max_{\hat{\pi} \in \mathcal{P}_{t-1}} f(\hat{\pi})$$

MAX, $\mathbb{E}$, Σ

b) FOR π, X_t IS A MARKOV CHAIN

IF $\pi_0(x) = a$ THEN
$$\mathbb{E}_{x,a}[Y] = \mathbb{E}_{x,a}\left[\mathbb{E}_{X_1,\pi_1}[Y]\right]$$

SEPARATE
PRESENT
& FUTURE.

c) $$\sum_{s=0}^{t-1} r(X_s,\pi_s) + r(X_t) = r(x,a) + \sum_{s=1}^{t-1} r(X_s,\pi_s) + r(X_t)$$

PROOF OF LEMMA

a) $$\pi = (\pi_0(x), \underbrace{\pi_1, \pi_2, \dots, \pi_{t-1}})$$

$$\mathcal{P}_t^n(x) = \mathcal{A}^n \times \mathcal{P}_{t-1}^n$$

$$\therefore \max_{\pi \in \mathcal{P}_t(x)} = \max_{a \in \mathcal{A}} \max_{\pi \in \mathcal{P}_{t-1}}$$

b) $$X_{t+1} = f(X_t, \pi_t(X_t); U_t)$$

BY MARKOV CHAIN PROP IS A MARKOV CHAIN.

$$\mathbb{E}_{x,\pi_0}[Y] = \mathbb{E}_{x,\pi_0}\left[\mathbb{E}[Y \mid X_1, A_1 = \pi_1(X_1), \cancel{X_0 = x, A_0 = \pi_0}]\right]$$

MARKOV

$$= \mathbb{E}_{x,\pi_0}\left[\mathbb{E}_{X_1, \pi(X_1)}[Y]\right]$$

c) TRIVIAL

$\square$

PROOF:

$$V_t(x) = \underbrace{\max_{\pi_0,\dots,\pi_{t-1}}}_{\max\limits_{\pi_0}\ \max\limits_{\pi_1,\dots,\pi_{t-1}}} \underbrace{\mathbb{E}_{x,\pi_0}}_{\mathbb{E}_{x,\pi_0}\ \mathbb{E}_{X_1,\pi_1}} \Bigg[\underbrace{\sum_{s=0}^{t-1} r(X_s,\pi_s) + r(X_t)}_{r(x,\pi_0) + \sum_{s=1}^{t-1} r(X_s,\pi_s) + r(X_T)} \Bigg]$$

MOVE AS FAR TO RIGHT AS POSSIBLE

$$= \max_{\underset{a}{\pi_0}} \left\{ r(x,\pi_0) + \mathbb{E}_{x,\pi_0}\Bigg[\underbrace{\max_{\pi_1,\dots,\pi_t} \mathbb{E}_{X_1,\pi_1}\Big[\sum_{s=1}^{t-1} r(X_s,\pi_s) + r(X_T) \Big]}_{V_{t-1}(\hat{X})} \Bigg] \right\}$$

$$= \max_{a} \left\{ r(x,a) + \mathbb{E}_{x,a}\big[V_{t-1}(\hat{X}) \big] \right\} \qquad \square$$

Ex 20. *You need to sell a car. At every time $t = 0, ..., T-1$, you set a price p_t and a customer then views the car. The probability that the customer buys a car at price p is $D(p)$. If the car isn't sold be time T then it is sold for fixed price V , V < 1. Maximize the reward from selling the car and find the recursion for the optimal reward when $D(p) = (1-p)_+$.*

SOLD OR NOT $\to x$, $p \to a$

$$V_t(x) = \begin{cases} V_t & \text{IF } x=0 \text{ NOT SOLD} \\ 0 & \text{IF } x=1 \text{ SOLD} \end{cases}$$

ANSWER:

$$V_t = \max_p \Big\{ \overbrace{p\,D(p)}^{r(x,a)} + (1-D(p))\,\overbrace{V_{t-1}} \Big\} \qquad V_0 = V$$

$$= \max_p \Big\{ p(1-p) + (1-(1-p))V_{t-1} \Big\}$$

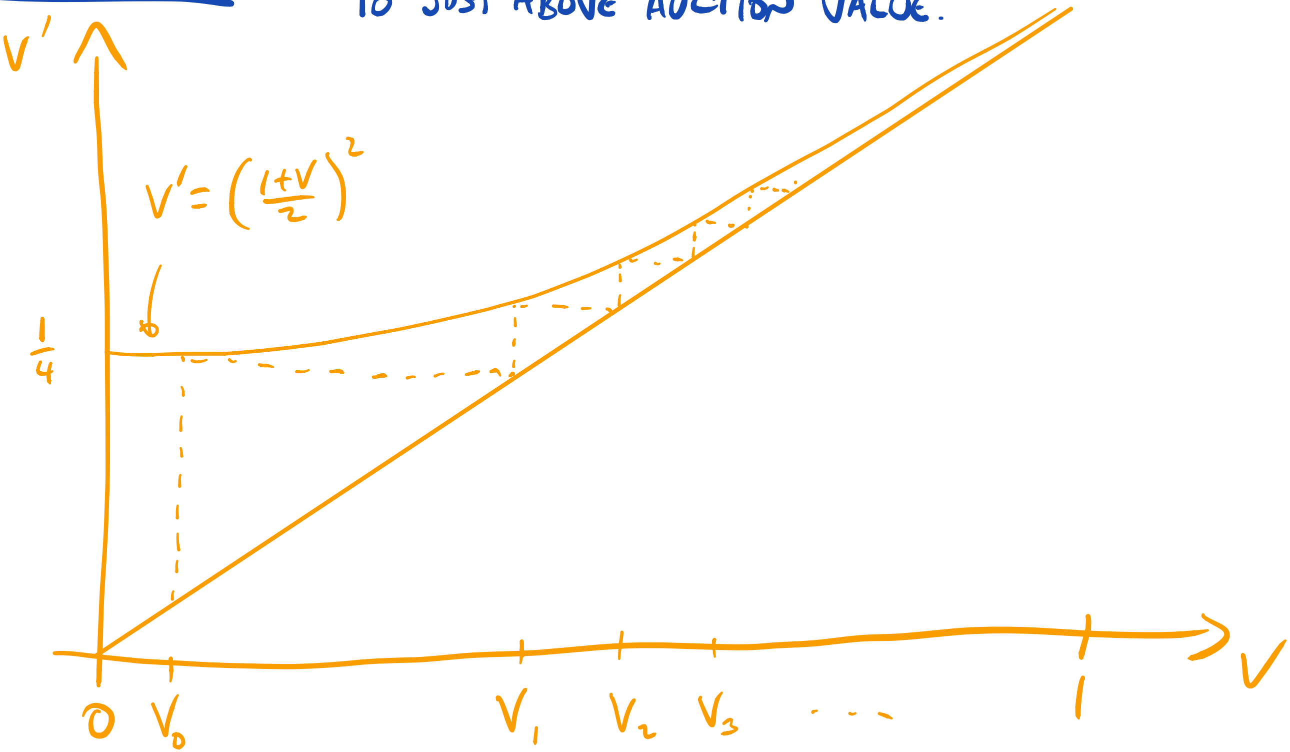
DIFFERENTIATE & SOLVE FOR p GIVES

$$V_t = \left(\frac{1+V_{t-1}}{2}\right)^2$$

NOTE $V_t > V_{t-1}$ & ONLY FIXED POINT IS $V^* = 1$

SO $V_t \to 1$ AS $t \to \infty$

A PICTURE: START HIGH & SLOWLY DECREASE PRICE TO JUST ABOVE AUCTION VALUE.



Ex 21 (Call Option). *You own a call option with strike price p. Here you can buy a share at price p making profit $X_t - p$ where x_t is the price of the share at time t. The share must be exercised by time T. The price of stock X_t satisfies*

$$X_{t+1} = X_t + \epsilon_t$$

for ϵ_t IIDRV with finite expectation. Show that there exists a decresing sequence $\{a_t\}_{0\leq t\leq T}$ such that it is optimal to exercise whenever $X_s \geq a_s$ occurs.

ANSWER: BELLMAN EQN IS

$$V_t(x) = \max\{x - p, \mathbb{E}[V_{t-1}(x+\epsilon)]$$

1). POLICIES WITH t STEPS LEFT CONTAIN ALL POLICIES WITH $t-1$ STEPS

SO $V_t(x) \geq V_{t-1}(x)$.

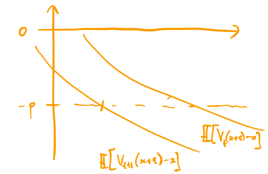
2) $V_t(x) - x = \max\{-p, \underbrace{\mathbb{E}[V_{t-1}(x-\epsilon) - x]}_{\text{DECREASING}}\}$ IS DECREASING CONTINUOUS IN x.

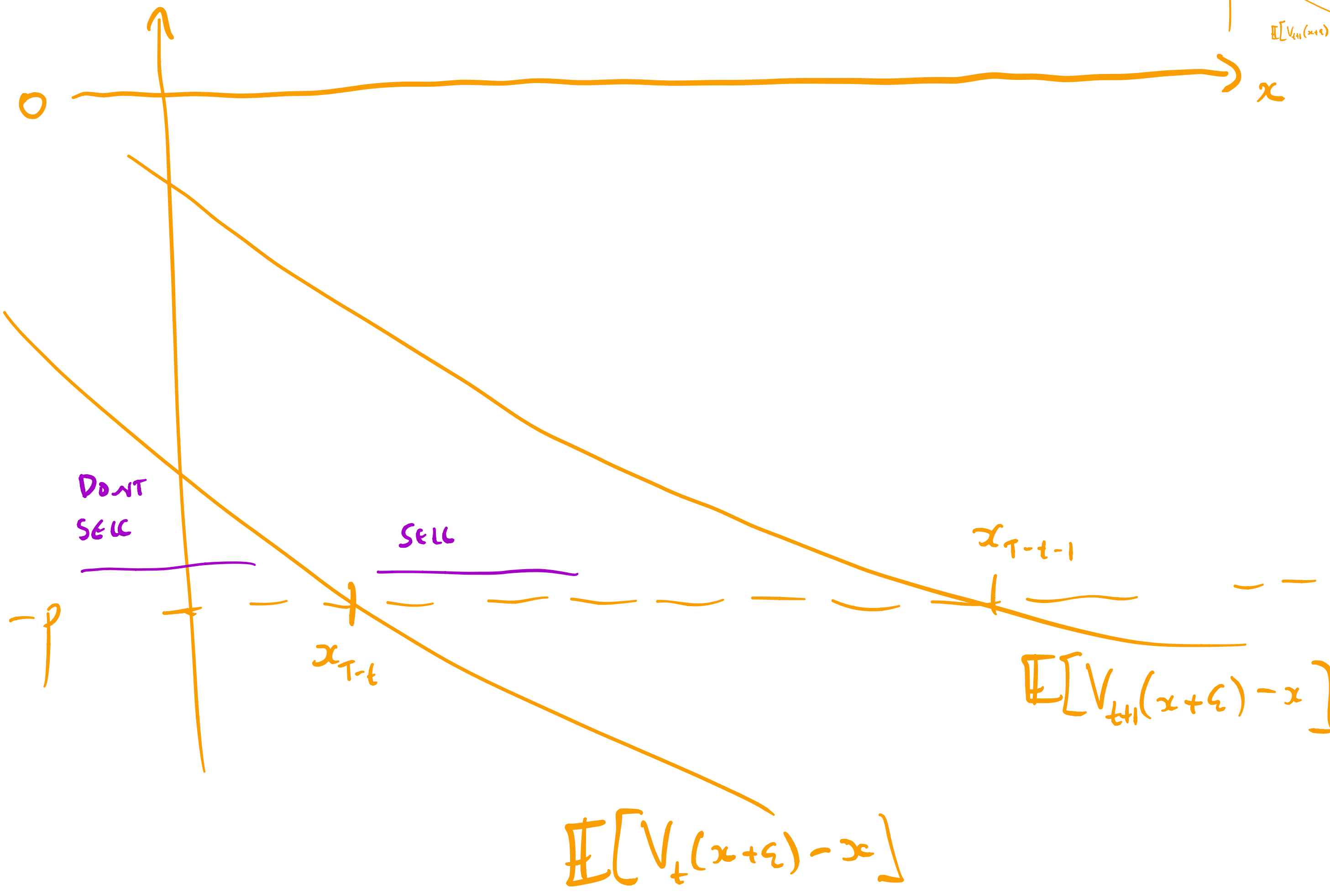
BECAUSE $V_0(x) - x = \max\{-p, -x\}$ HAS THESE PROPERTIES. & MAX & EXPECTATION PRESERVES THESE PROPERTIES.

$\therefore$ DECISION SWITCHES AT x_{T-t} s.t. $-p = \mathbb{E}[V_{t-1}(x_{T-t} - \epsilon) - x_{T-t}]$

x_t DECREASES BECAUSE V_t INCREASES

A PICTURE:


0
-p
$\mathbb{E}[V_t(x+\varepsilon)-x]$
$\mathbb{E}[V_{t+1}(x+\varepsilon)-x]$


0
x
DONT SELL
SELL
$-p$
x_{T-t}
x_{T-t-1}
$\mathbb{E}[V_{t+1}(x+\varepsilon)-x]$
$\mathbb{E}[V_t(x+\varepsilon)-x]$