4. INFINITE TIME HORIZON

[REALLY INDEFINATE TIME LENGTH.]

INFINITE TIME

INFINITE

DISCOUNT FACTOR $\beta \in (0,1]$

$$V(x) = \underbrace{\max_{\pi}}_{\max\limits_{\pi_0} \max\limits_{\pi_1,\pi_2,\dots}} \underbrace{\mathbb{E}_x}_{\mathbb{E}_x \mathbb{E}_{X_1}} \Big[\underbrace{\sum_{t=0}^{\infty} \beta^t r(X_t)}_{r(x) + \beta \sum_{t=1}^{\infty} \beta^{t-1} r(X_t)} \Big]$$

$$= \max_{\pi_0} \mathbb{E}_{x,\pi_0}\Big[r(x) + \beta\, \underbrace{\mathbb{E}_{X_1}\Big[\max_{\pi_1,\pi_2,\dots} \sum_{t=1}^{\infty} \beta^{t-1}(X_t) \Big]}_{\mathbb{E}_{X_1} V(X_1)} \Big]$$

$$= \max_{\pi_0} r(x) + \beta\, \mathbb{E}_{x,\pi_0}\big[V(X_1) \big] \quad \leftarrow \text{BELLMAN EQUATION}$$

[BEFORE SOLVING THIS GAVE THE OPTIMAL POLICY, BUT WHAT ABOUT NOW?]

DISCOUNTED, POSITIVE, NEGATIVE & AVERAGE PROGRAMMING

$$V(x) = \max_{\pi} \lim_{T\to\infty} \mathbb{E}_x \left[\sum_{t=0}^{T} \beta^t r(X_t, \pi_t) \right]$$

LIMIT & MAX INTERACT. SO DIFFERENT CASES

- DISCOUNTED PROGRAMMING: $\beta \in (0,1)$, $\max_{x,a} |r(x,a)| < \infty$
- POSITIVE PROGRAMMING: $\beta \in (0,1]$, $r(x,a) \geq 0$
- NEGATIVE PROGRAMMING: $\beta \in (0,1]$, $r(x,a) \leq 0$
  [OR MINIMIZE $c(x,a) \geq 0$]
- AVERAGE PROGRAMMING:

$$\lim_{T\to\infty} \frac{1}{T} \mathbb{E}_x \left[\sum_{t=0}^{T} r(X_t, \pi_t) \right]$$

DISCOUNTED PROGRAMMING RESULT.

Thrm 43. *For a discounted program, the optimal policy $V(x)$ satisfies*

$$V(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[V(\hat{X}) \right] \right\}.$$

Moreover, if we find a function $R(x)$ such that

$$R(x) = \max_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[R(\hat{X}) \right] \right\}$$

then $R(x) = V(x)$, i.e. the solution to the Bellman equation is unique, and we find a function $\pi(x)$ such that

$$\pi(x) \in \operatorname*{argmax}_{a\in\mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[R(\hat{X}) \right] \right\}$$

Then π is optimal and $R(x,\pi) = R(x) = V(x)$ the optimal value function.

PROOF: FIRST LETS SHOW THE BELLMAN EQUATION HOLDS:

$$R_t(x,\pi) = r(x,\pi_0) + \beta\,\mathbb{E}\left[R_{t-1}(\hat{x},\hat{\pi})\right]$$

TAKE LIMITS $t \to \infty$

$$R(x,\pi) = r(x,\pi_0) + \beta\,\mathbb{E}_{x,\pi_0}\left[R(\hat{x},\hat{\pi})\right]$$

$$\leq r(x,\pi_0) + \beta\,\mathbb{E}_{x,\pi_0}\left[V(\hat{x})\right]$$

NOW MAXIMIZE OVER π_0 & π,

$$V(x) = \max_{\pi} R(x,\pi)$$

$$\leq \max_{a}\left\{r(x,a) + \beta\,\mathbb{E}_{x,a}\left[V(\hat{x})\right]\right\}$$

SO WE HAVE

$$V(x) \leq \max_{a}\left\{r(x,a) + \beta\,\mathbb{E}_{x,a}\left[V(\hat{x})\right]\right\}.$$

LET $\hat{\pi}$ BE SUCH THAT $R(\hat{x}, \hat{\pi}) \geq V(\hat{x}) - \varepsilon$

THEN FOR POLICY π THAT DOES a THEN $\hat{\pi}$

$$
\begin{aligned}
V(x) &\geq R(x, \pi) \\
&= r(x,a) + \beta \, \mathbb{E}_{x,a}[R(\hat{x}, \hat{\pi})] \\
&\geq r(x,a) + \beta \, \mathbb{E}_{x,a}[V(\hat{x})] - \beta \varepsilon
\end{aligned}
$$

LET $\varepsilon \to 0$ & THEN MAXIMIZE OVER a:

$$V(x) \geq \max_a \{ r(x,a) + \beta \, \mathbb{E}_{x,a}[V(\hat{x})]]$$

$\therefore$ BELLMAN HOLDS

$$V(x) = \max_a \{ r(x,a) + \beta \, \mathbb{E}_{x,a}[V(\hat{x})] \} \quad \checkmark$$

PROOF (CONTINUED):

WANT: TO SHOW UNIQUENESS OF SOLN

[SIMILAR TO UNIQNESS FOR MARKOV CHAINS].

FIRST A DEFINITION

Def 9 (Q-Factor). *The Q-factor of reward function $R(\cdot)$ is the value for taking action a in state x and then at the next step receiving reward $R(\hat{X})$:*

$$Q_R(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta R(\hat{X})]\,.$$

Similarly the Q-factor for a policy π, denoted by $Q_\pi(x,a)$, is given by the above expression with $R(x) = R(x,\pi)$. The Q-factor of the optimal policy is given by

$$Q^*(x,a) = \max_\pi Q_\pi(x,a).$$

SUPPOSE $R(x)$ IS ANOTHER SOL$\underline{^N}$. SO $R(x) = \max_a Q_R(x,a)$

THEN

$$Q_V(x,a) - Q_R(x,a) = \beta \mathbb{E}_{x,a}\left[V(\hat{x}) - R(\hat{x})\right]$$

$$= \beta \mathbb{E}_x\left[\max_a Q(\hat{x},a) - \max_{a'} Q(\hat{x},a')\right]$$

SO

$$\|Q_V - Q_R\|_\infty \leq \beta \max_x \left| \max_a Q_V(\hat{x},a) - \max_{a'} Q_R(\hat{x},a') \right|$$

$$\leq \beta \max_{x,a} \left| Q_V(\hat{x},a) - Q_R(\hat{x},a') \right|$$

$$= \beta \|Q_V - Q_R\|_\infty$$

$\therefore$ ONLY SOLUTION IS $Q_V = Q_R$

$\therefore$ $R(x) = \max_a Q_R(x,a) = \max_a Q_V(x,a) = V(x)$ ✓ SOLUTION IS UNIQUE

PROOF (CONTINUED):

WANT: IF $\pi(x) \in \underset{a}{\mathrm{ARGMAX}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a}\left[R(\hat{X})\right] \right\}$

THEN $R(x,\pi) = R(x)$. $(+)$.

TO SHOW THIS NOTE $R(x)$ SOLVES

$$R(x) = r(x,\pi(x)) + \beta \mathbb{E}_{x,a}\left[R(\hat{x})\right] \qquad (\#)$$

WHERE UNDER π X_t IS NOW A MARKOV CHAIN.

BY OUR PROPOSITION ON MARKOV CHAINS

SOLUTION TO $(\#)$ IS UNIQUE & GIVEN BY $R(x,\pi)$

SO $R(x) = R(x,\pi)$ $\therefore \pi$ IS OPTIMAL. $\square$

Some facts on Q-factors:

Prop 49. *a) Stationary Q-factors satisfy the recursion*

$$Q_\pi(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta Q_\pi(\hat{X}, \pi(\hat{X}))] \,.$$

b) Bellman's Equation can be re-expressed in terms of Q-factors as follows

$$Q^*(x,a) = \mathbb{E}_{x,a}[r(x,a) + \beta \max_{\hat{a}} Q^*(\hat{X}, \hat{a}))] \,.$$

The optimal value function satisfies

$$V(x) = \max_{a \in \mathcal{A}} Q^*(x,a).$$

c) The operation

$$F_{x,a}(\boldsymbol{Q}) = \mathbb{E}_{x,a}[r(x,a) + \beta Q_\pi(\hat{X}, \pi(\hat{X}))]$$

is a contraction with respect to the supremum norm, that is,

$$\|\boldsymbol{F}(\boldsymbol{Q}_1) - \boldsymbol{F}(\boldsymbol{Q}_2)\|_\infty \leq \|\boldsymbol{Q}_1 - \boldsymbol{Q}_2\|_\infty \,.$$

POSITIVE PROGRAMMING.

Thrm 50. *Consider a positive program the optimal value function $V(x)$ is the minimal non-negative solution to the Bellman equation*

$$R(x) = \max_{a \in \mathcal{A}} \left\{ r(x,a) + \beta \mathbb{E}_{x,a} \left[R(\hat{X}) \right] \right\}.$$

Thus if we find a policy π whose reward function $R(x,\pi)$ satisfies the Bellman equation. Then it is optimal.

PROOF: SOLUTION OVER T+1 STEPS IS

$$V_{T+1}(x) = \max_a \; r(x,a) + \beta \mathbb{E}_{x,a} [V_T(\hat{x})]$$

WITH $V_0(x) = 0$. $V_T(x)$ IS INCREASING IN T

LET

$$V_\infty(x) = \sup_T V_T(x) = \lim_{T \to \infty} V_T(x)$$

$$V_\infty(x) = \sup_T \max_{a\in\mathcal{A}} r(x,a) + \beta\,\mathbb{E}_{x,a}\left[V_T(\hat{X})\right]$$

MONOTONE CONVERG.

$$= \max_a r(x,a) + \beta\,\mathbb{E}_{x,a}\left[\sup_T V_T(\hat{X})\right]$$

$$= \max_a r(x,a) + \beta\,\mathbb{E}_{x,a}\left[V_\infty(\hat{x})\right]$$

CLEARLY $V(x) \geq V_T(x)$ $\therefore$ $V(x) \geq V_\infty(x)$

BUT

$$V_T(x) \geq R_T(x,\pi) \;\; \forall \pi \quad \therefore \quad V_\infty(x) \geq R(x,\pi) \;\; \forall \pi$$

$$\therefore \quad V_\infty(x) \geq V(x)$$

So $V_\infty(x) = V(x)$. $\square$ [THIS GIVES VALUE ITERATION ARGUMENT].

NEGATIVE PROGRAMMING:

NEEDS STRONGER CONDITIONS TO WORK

$$L_T(x) = \min_{\pi} C_T(x,\pi) \le L(x)$$

i.e. $$\max_T \min_{\pi} C_T(x,\pi) \le \min_{\pi} \max_T C_T(x,\pi)$$

[NOT OBVIOUS HOW TO SWAP MIN & MAX]

Thrm 52. *Consider a negative program, minimizing positive costs. For the minimal non-negative solution to the Bellman equation*

$$L(x) = \min_{a \in \mathcal{A}} \left\{ l(x,a) + \beta \mathbb{E}_{x,a}\left[L(\hat{X})\right] \right\}, \tag{1.8}$$

any stationary policy Π *that solves the Bellman equation:*

$$\pi(x) \in \operatorname*{argmin}_{a \in \mathcal{A}} \left\{ c(x,a) + \beta \mathbb{E}_{x,a}\left[L(\hat{X})\right] \right\}$$

is optimal.

PROOF: USE ROLL OUT.

$$
\begin{aligned}
L(x) &= \min_{a \in \mathcal{A}} \{c(x,a) + \beta \mathbb{E}_{x,a}[L(X_1)]\} \\
&= c(x, \pi(x)) + \beta \mathbb{E}_{x,\pi(x)}\,[L(X_1)] \\
&= c(X_0, \pi(X_0)) + \beta \mathbb{E}_{X_0,\pi(X_0)} \Big[c(X_1, \pi(X_1)) + \beta \mathbb{E}_{X_1,\pi(X_1)}\,[L(X_2)]\Big] \\
&= C_1(x,\pi) + \beta^2 \mathbb{E}_{x,\pi}[L(X_2)] \\
&\;\vdots \\
&= C_T(x,\pi) + \beta^T \mathbb{E}_{x,\pi}[L(X_{T+1})]\,.
\end{aligned}
$$

Thus

$$
L(x) = C_T(x,\pi) + \beta^T \mathbb{E}_{x,\pi}[L(x_{T+1})] \geq C_T(x,\pi) \xrightarrow[T\to\infty]{M.C.T.} C(x,\pi)\,.
$$

So the policy has lower cost, and thus is optimal.

AVERAGE PROGRAMMING:

CONSIDER

$$\bar{R}(\pi) = \lim_{T \to \infty} \frac{1}{T} \overbrace{\mathbb{E}_x\left[\sum_{t=0}^{T} r(X_t, \pi_t)\right]}^{R_T(x,\pi)}$$

WANT

$$\bar{V} = \max_{\pi} \bar{R}(\pi)$$

[NOTICE $\bar{V}$ NO LONGER DEPENDS ON x
BECAUSE X_t SHOULD BE CONVERGING TO AN EQUILIBRIUM.]

A BELLMAN EQUATION STILL HOLDS

THINK:

$$R_T(x,\pi) = \underbrace{T \cdot V_\pi}_{\text{LONG RUN VALUE OF } \pi} + \underbrace{\rho(x)}_{\text{SHORT TERM EFFECT FOR BEING IN } x.} + \underbrace{\varepsilon(x,T)}_{\text{OTHER VANISHING TERM}}$$

$\varepsilon(x,T) \to 0$

AVERAGE PROGRAMMING:

FOR $V_T(x) = T \cdot V^* + \rho(x)$

THE BELLMAN EQUATION BECOMES:

$$V_{T+1}(x) = \max_a \left\{ r(x,a) + \mathbb{E}_{x,a}\left[V_T(\hat{x}) \right] \right\}$$

$$V_{T+1}(x) = (\cancel{T}+1)V^* + \rho(x)$$

$$V_T(\hat{x}) = \cancel{T}V^* + \rho(\hat{x})$$

$$\therefore \quad V^* = \max_a \left\{ r(x,a) + \mathbb{E}_{x,a}\left[\rho(\hat{x}) \right] - \rho(x) \right\}$$

AVERAGE PROGRAMMING:

THE SAME RESULT HOLDS FOR MINIMIZING

THEOREM: IF WE CAN FIND A CONSTANT $\bar{V}$

& A BOUNDED FUNCTION $\rho(x)$ SUCH THAT $\forall x$

$$V^* = \max_a r(x,a) + \mathbb{E}_{x,a}\left[\rho(\hat{x})\right] - \rho(x)$$

& IF GIVEN THIS V^* & $\rho(x)$ WE CAN CHOSE A POLICY SUCH THAT

$$\pi^*(x) \in \operatorname{argmax}_a r(x,a) + \mathbb{E}_{x,a}\left[\rho(\hat{x})\right] - \rho(x)$$

THEN

V^* IS THE MAXIMUM AVERAGE REWARD

& $\pi^*(x)$ IS THE OPTIMAL POLICY

PROOF: WE PROVE THE RESULT BY PROVING INEQUALITIES IN BOTH DIRECTIONS.

IF V, p IS SUCH THAT $\quad V \geq \max_x \max_a r(x,a) + \mathbb{E}_{x,a}[p(\hat{x})] - p(x) \quad (+)$

THEN, WE CLAIM THAT,

$$V \geq \limsup_{T\to\infty} \frac{R_T(x,\pi)}{T} \qquad \forall \text{ POLICIES } \pi$$

PROOF OF CLAIM: $M_0 = 0$

DEFINE $\quad M_t - M_{t-1} = r(X_t, \pi_t) - V + \mathbb{E}_{X_t \pi_t}[p(\hat{x})] - p(X_t)$

CLEARLY $\quad \mathbb{E}[M_t - M_{t-1} | \mathcal{F}_{t-1}] \leq 0 \quad \therefore M_t$ IS A SUPER-M_g BY (+)

$$\therefore 0 \geq \frac{\mathbb{E}_x[M_T]}{T} = \bar{R}_T(x,\pi) - V + \frac{\mathbb{E}_x[p(X_{T-1})]}{T} - \frac{p(x)}{T}$$

LETTING $T \to \infty \qquad V \geq \bar{R}(x,\pi) \qquad \forall \pi$

FROM THIS CLAIM WE SEE THAT $\quad V^* \geq \bar{R}(x,\pi)$

PROOF (CONT)

SUPPOSE THAT $\pi^*(x)$, IS THE POLICY AS DESCRIBED IN THE THEOREM

$$M_t^* - M_{t-1}^* = r(X_t, \pi_t) - V^* + \mathbb{E}_{X_t \pi_t}[p(\hat{x})] - p(X_t)$$

$\therefore$ M_t^* IS A MARTINGALE &

$$0 = \frac{\mathbb{E}_x[M_T^*]}{T} = \bar{R}_T(x, \pi^*) - V^* + \mathbb{E}_x\left[\frac{p(X_{T+1})}{T}\right] - \frac{p(x)}{T}$$

$$\therefore \lim_{T \to \infty} \bar{R}_T(x, \pi^*) = V^* \underset{\uparrow}{\geq} \limsup_{T \to \infty} \bar{R}_T(x, \pi)$$

BY FIRST PRN

$\therefore$ π^* IS OPTIMAL & ITS VALUE IS V^* $\square$.

MARTINGALE PRINCIPLE OF OPTIMALITY.

Prop 54 (A Martingale Principle of Optimal Control.). *Consider discounted program. Suppose for a bounded function $R : \mathcal{X} \to \mathbb{R}$ we define a process $(M_t : t \in \mathbb{Z}_+)$ whose increments, $\Delta M(X_t) := M_{t+1} - M_t$, are given by*

$$\Delta M(x) = R(x) - \beta R(\hat{x}) - r(x, \pi(x))$$

If M_t is a supermartingale for all policies π' and, for some π, M_t is a martingale, then π is the optimal policy and $R(x) = R(x, \pi)$.

Proof. M_t is a martingale [resp. supermartingale] iff

$$M_t^\beta := \sum_{s=0}^{\infty} \beta^s \Delta M(X_s)$$

is a martingale [resp. supermartingale]. Taking expecations,

$$0 \leq \mathbb{E}_x[M_t^\beta] = \mathbb{E}_x\Big[R(x) - \beta^{t+1} R(X_{t+1}) - \sum_{s=0}^{t} \beta^s r(X_s, \pi(X_s))\Big]$$

Rearranging and letting $t \to \infty$ gives, for π',

$$R(x) \geq \mathbb{E}\Big[\sum_{s=0}^{\infty} \beta^s r(X_s, \pi'(X_s))\Big],$$

where the inequality above holds with equality if M_t^β is a martingale for some π . Thus we see that $R(x) \geq V(x)$, where $V(x)$ is the value function for the MDP and $R(x) = V(x) = R(x, \pi)$. □

SUMMARY: INFINITE TIME MDPs

- BELLMAN'S EQN STILL HOLDS

$$V(x) = \max_{a} \left\{ r(x,a) + \beta \, \mathbb{E}_{x,a}\left[V(\hat{x}) \right] \right\}$$

OR

$$Q(x,a) = r(x,a) + \beta \, \mathbb{E}_{x,a}\left[\max_{a'} Q(\hat{x},a) \right]$$

- DISCOUNTED PROGRAMMING $\beta \in (0,1)$, $\|r\|_\infty < \infty$
  - ANY SOLUTION TO BELLMAN WILL DO
  - PROOF USES CONTRACTION PROPERTY OF $Q(x,a)$

- POSITIVE PROGRAMMING $r \geq 0, \quad \beta = 1$
    - MINIMAL SOLUTION TO BELLMAN OR A POLICY WILL DO
    - PROOF REPEATEDLY SOLVE FINITE TIME BELLMAN EQN.
- NEGATIVE PROGRAM $c \geq 0$ OR $r \leq 0, \quad \beta = 1$
    - MINIMAL SOLUTION TO BELLMAN & A STATIONARY POLICY!
    - PROOF ROLL OUT BELLMAN EQN.
- AVERAGE PROGRAM $\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^{T} r(x_t, a_t)\right]$
    - FIND k, & λ s.t. $k(x) = \min_a \left\{ c(x,a) - \lambda + \mathbb{E}_{x,a}[k(\hat{x})] \right\}$
    - PROOF MARTINGALE ARGUMENT.