Linear Programming (LP) for MDPs.

Primal form. $\min_{V \in \mathbb{R}^S} \; d_0^T V$ → "init state distribution" $d_0 > 0$, $\sum_s d_0(s) = 1$.

s.t. $V \geq \mathcal{T} V$ → Bellman optimality op.

① Why LP? Constraints: $\forall s \in S$.

$$V(s) \geq \max_{a \in A} \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V(s')] \right).$$

Nonlinear! Can convert to $|A|$ linear constraints:

$$V(s) \geq R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V(s')], \quad \forall a \in A.$$

linear!

∴ Primal form has $|S|$ decision varibles, & $|S| \times |A|$ constraints.

② Why it solves planning poblem?

Claim: if $d_0(s) > 0, \forall s \in S$. then opt. sol. to poimal is $V = V^*$.

Proof: constraint $V \geq \mathcal{T} V$.

Lemma. monotone property of $\mathcal{T}$. $\forall V \geq V'$. $\mathcal{T}V \geq \mathcal{T}V'$.

We will show: $V \geq \mathcal{T}V \Rightarrow V \geq V^*$.

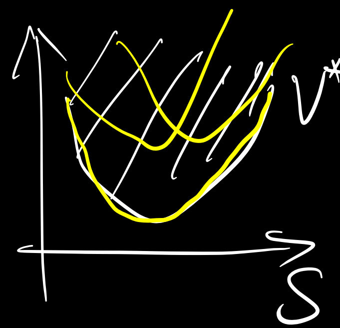
Invoke monotonicity on V & $V' = \mathcal{T}V$. $\Rightarrow \mathcal{T}V \geq \mathcal{T}(\mathcal{T}V)$

$V \geq \mathcal{T}V \geq \mathcal{T}^2 V \Rightarrow V \geq \mathcal{T}^2 V$.

Now let $V' = \mathcal{T}^2 V \Rightarrow V \geq \mathcal{T}V \geq \mathcal{T}^3 V \Rightarrow V \geq \mathcal{T}^3 V$.

… $\forall n$, $V \geq \mathcal{T}^n V$.

take limit $n \to \infty$. $\quad V \geq \mathcal{T}^{\infty} V = V^*$.



Therefore ① any feasible $V \geq V^*$.

② V^* is feasible: $V^* = \mathcal{T} V^*$

$\Rightarrow$ $\min d_0^T V$ is achieved w/ $V = V^*$.

$\boxed{\text{Dual form}}$

$$\max_{d \in \mathbb{R}^{S \times A},\, d \geq 0} \quad d^T R \longrightarrow \text{reward function.}$$

$$\text{s.t. } \forall s \in S, \quad \sum_a d(s,a) = \gamma \sum_{\tilde{s},\tilde{a}} d(\tilde{s},\tilde{a}) P(s|\tilde{s},\tilde{a}) + (1-\gamma) d_0(s).$$

Interpretation: d plays the role of occupancy.

In fact. dual constraint characterizes all possible occupancies in the MDP.

if $d = d^{\pi}$. $\Rightarrow$ $d^T R = (d^{\pi})^T R = \mathbb{E}_{(s,a) \sim d^{\pi}} [R(s,a)]$.

$$= \mathbb{E}_{s \sim d_0} [V^{\pi}(s)].$$

Verify $d = d^{\pi}$ is feasible for any π:

$$d^{\pi}(s,a) = (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi}(s,a)$$

distribution of s_t, a_t under π, d_0.

$$\underline{d^{\pi}(s)} = (1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \underline{d_t^{\pi}(s)}$$

Dual constraint: $\displaystyle \sum_a d(s,a) = \gamma \sum_{\tilde{s},\tilde{a}} d(\tilde{s},\tilde{a}) P(s|\tilde{s},\tilde{a}) + (1-\gamma) d_0(s)$

$$\text{LHS} = \sum_{a \in A} d^{\pi}(s,a) = d^{\pi}(s) = \underline{(1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1} d_t^{\pi}(s)}.$$

$$\text{RHS (first term)} = \gamma(1-\gamma)\sum_{t=1}^{\infty} \underbrace{\sum_{\tilde{s},\tilde{a}} \gamma^{t-1} d_t^{\pi}(\tilde{s},\tilde{a})\, P(s|\tilde{s},\tilde{a})}_{\text{a function } s.}$$

a distribution of state.

↳ s drawn from the dist $\Leftrightarrow (\tilde{s},\tilde{a}) \sim d_t^{\pi}$, $s \sim P(\cdot|\tilde{s},\tilde{a})$.

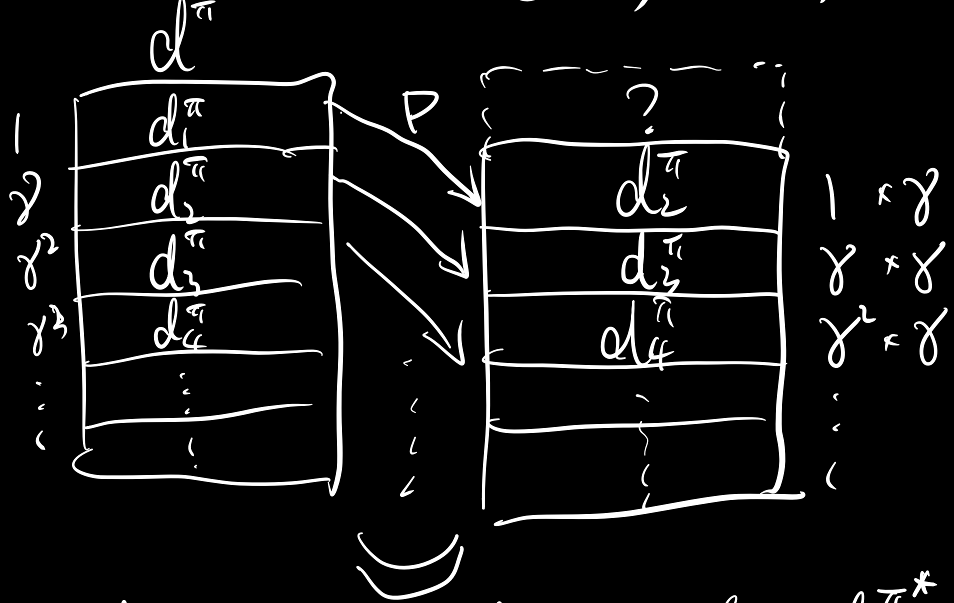
$\Rightarrow$ the distribution is $d_{t+1}^{\pi}(s)$.

$$\Rightarrow \text{RHS (first term)} = \gamma(1-\gamma)\sum_{t=1}^{\infty} \gamma^{t-1} d_{t+1}^{\pi}(s)$$

$$= (1-\gamma)\sum_{t=1}^{\infty} \gamma^{t} d_{t+1}^{\pi}(s)$$

$$= (1-\gamma)\sum_{t=2}^{\infty} \gamma^{t-1} d_{t}^{\pi}(s)$$

$$\text{RHS (second term)} = (1-\gamma)\, d_0(s) = (1-\gamma)\, d_1^{\pi}(s) \quad \forall \pi.$$



$$\boxed{d^{\pi}(s,a) = d^{\pi}(s) \cdot \pi(a|s)}$$

Remark: solution to dual is $d = d^{\pi^*} \xrightarrow{\text{back out}} \pi^*$.

$\forall \pi$ (stat. possibly stochastic), given $d^{\pi}(s,a) \Rightarrow \pi(a|s) = \dfrac{d^{\pi}(s,a)}{\sum_{a'} d^{\pi}(s,a')}$