MDP with finite state & action:

setting: $a \in A$, $s \in S$, $P^a_{st}$ ($\sum_t P^a_{st} = 1 \ \forall s, a$)

($P^a$ is a transition matrix)

$\gamma \in (0,1)$, $r^a_s$, $\pi^a_s$ ($\sum_a \pi^a_s = 1 \ \forall s$)

Define $V^\pi_s = \underset{\substack{\downarrow \\ a_t \sim \pi^\cdot_{s} \\ s_{t+1} \sim P^{a_t}_{s_t,\cdot}}}{\mathbb{E}} \left[ \sum_{t=0}^{\infty} \gamma^t r^{a_t}_{s_t} \mid s_0 = s \right]$

$V^*_s = \max_\pi V^\pi_s$ & $V^* = \begin{pmatrix} V^*_1 \\ \vdots \\ V^*_{|S|} \end{pmatrix}$.

Bellman Eq: $V^\pi_s = r^\pi_s + \gamma \sum_t P^\pi_{st} V^\pi_t$ $\quad r^\pi_s = \sum r^a_s \pi^a_s$, $P^\pi_{st} = \sum_a P^a_{st} \pi^a_s$

pf: $V^\pi_s = \gamma \sum_a r^a_s \pi^a_s + \gamma \underbrace{\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r^{a_{t+1}}_{s_{t+1}} \mid s_0 = s\right]}$

$= \sum_a \mathbb{E}[V_{s_1} \mid s_0 = s, a_0 = a] \pi^a_s = \sum_{s_1, a} P^a_{ss_1} V_{s_1} \pi^a_s$

optimal Bellman Eq: $V^*_s = \max_a \left( r^a_s + \gamma \sum_t P^a_{st} V^*_t \right)$

$\Updownarrow$ ①

(P) $\begin{cases} \min_{V \in \mathbb{R}^{|S|}} \sum e_s V_s \\ \text{s.t.} \quad r^a_s + \gamma \sum P^a_{st} V_t - V_s \le 0 \quad \forall a, s. \end{cases}$

$\Updownarrow$ ②

(D) $\begin{cases} \max_{\lambda \in \mathbb{R}^{|A| \times |S|}} \sum_a (\lambda^a)^T r^a \\ \text{s.t.} \quad \lambda^a_s \ge 0 \quad \forall a, s \\ \qquad \sum_a (I - \gamma P^a)^T \lambda^a = e \end{cases}$

$\rightarrow$ $|S| \times |A|$-dim max prob with $|A| \times |S| + |S|$ constraints.

$\Updownarrow$ ③

$$e^T V^* = \max_{\pi \in \Delta^{|S|}} e^T V^\pi$$

s.t. $\begin{cases} V^\pi = r^\pi + \gamma P^\pi V \\ \sum_a \pi_s^a = 1, \quad \pi_s^a \geq 0 \end{cases}$ $\quad \left( P^\pi_{st} = \sum_a P^a_{st} \pi^a_s \right)$

(P) $V^* = \min \ e^T V$ $\qquad$ $|S|$-dimensional min prob.

$$\begin{cases} r^{a_1} + \gamma P^{a_1} V - V \leq 0 \\ \vdots \\ r^{a_m} + \gamma P^{a_m} V - V \leq 0 \end{cases}$$

$|A| \times |S|$ ineq constraints.

MDP with 2 actions, 2 states $\quad A = \{1, 2\}$, $\cancel{S} = \{1, 2\}$

$$P^1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad P^2 = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

$$r^1 = \begin{pmatrix} 1 \\ \frac{1}{2} \end{pmatrix}, \quad r^2 = \begin{pmatrix} \frac{3}{4} \\ \frac{1}{4} \end{pmatrix}, \quad \gamma = \frac{1}{2}$$

(P) : $\min \ \frac{1}{2}(V_1 + V_2)$

s.t.
$$1 + \frac{1}{2} V_2 \leq V_1$$
$$\frac{1}{2} + \frac{1}{2} V_1 \leq V_2$$
$$\frac{3}{4} + \frac{1}{2}\left(\frac{1}{2} V_1 + \frac{1}{2} V_2\right) \leq V_1$$
$$\frac{1}{4} + \frac{1}{2}\left(\frac{1}{2} V_1 + \frac{1}{2} V_2\right) \leq V_2$$

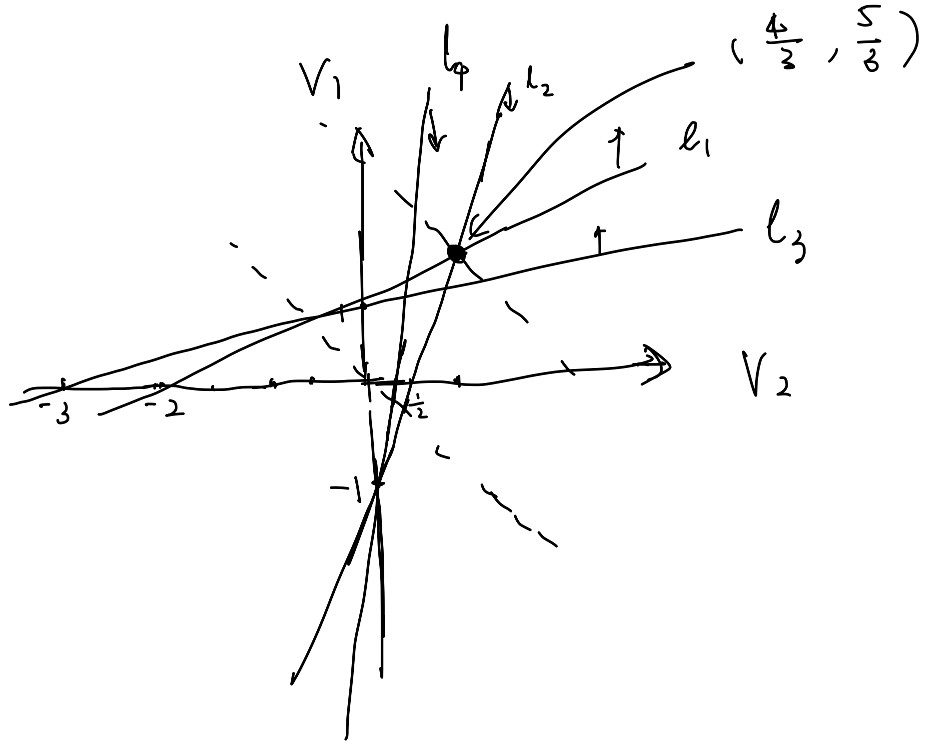
$\Rightarrow$

$V_1 \geq \frac{1}{2} V_2 + 1 \leftarrow \ell_1$

$V_1 \leq 2V_2 - 1 \leftarrow \ell_2$

$V_1 \geq \frac{1}{3} V_2 + 1 \leftarrow \ell_3$

$V_1 \leq 3V_2 - 1$




$$(D): V^* = \max \quad \lambda^1_1 + \frac{1}{2}\lambda^1_2 + \frac{3}{4}\lambda^2_1 + \frac{1}{4}\lambda^2_2$$

$$\text{s.t.} \quad \lambda^1_1, \lambda^1_2, \lambda^2_1, \lambda^2_2 \geq 0$$

$$\underset{a}{\lambda^1_1} + \underset{c}{\lambda^2_1} - \frac{1}{2}\left(\lambda^1_2 + \frac{1}{2}\lambda^2_1 + \frac{1}{2}\lambda^2_2\right) = \frac{1}{2}$$

$$\underset{b}{\lambda^1_2} + \underset{d}{\lambda^2_2} - \frac{1}{2}\left(\lambda^1_1 + \frac{1}{2}\lambda^2_1 + \frac{1}{2}\lambda^2_2\right) = \frac{1}{2}$$

$$\Leftrightarrow \quad \max \quad 2a + \frac{3}{2}c - \frac{1}{2} = k$$

$$\text{s.t.} \quad a, b, c \geq 0$$

$$4a + 3c - 2b \geq 2$$

$$5a + 4c - b = 4$$

$$\Rightarrow \quad a = b = 1, \quad c = d = 0.$$

$$\Rightarrow \begin{cases} \lambda^1_1 = 1 & \lambda^2_1 = 0 \\ \lambda^1_2 = 1 & \lambda^2_0 = 0 \end{cases}$$ $\Rightarrow$ always ~~take~~ action 1 for both states

$$\Rightarrow \begin{cases} V_1^* = \frac{5}{3}, \quad V_2^* = \frac{4}{3}. \\ \text{at both states, optimal action are } 1 \end{cases}$$

---

③ $\pi_s^a = \frac{1}{w_s} \lambda_s^a$ ← normalized $\lambda$, $w_s = \sum_a \lambda_s^a$

1) then $\lambda_s^a \geq 0 \ \forall a, s \iff \begin{cases} \pi_s^a \geq 0 \ \forall a, s \\ \sum_a \pi_s^a = 1 \end{cases} \iff \pi \in \Delta^{|S|}$

2) $\sum_a (I - \gamma P^a)^T \lambda^a = e \iff w = (I - \gamma P^\pi)^{-T} e$

3) $\sum_a (\lambda^a)^T r^a \iff w^T r^\pi$

4) $\max_{\pi \in \Delta^{|S|}} e^T (I - \gamma P^\pi)^{-1} r^\pi \iff \max_{\pi \in \Delta^{|S|}} e^T V^\pi$

s.t. $V^\pi = (I - \gamma P^\pi)^{-1} r^\pi$

$\Updownarrow$

$V^\pi = r^\pi - \gamma P^\pi V^\pi$

---

② Lagrangian $L(V, \lambda) = \sum_s e_s V_s + \sum_{a,s} \lambda_s^a (r_s^a + \gamma \sum_t P_{st}^a V_t - V_s)$

$\min_V L(V, \lambda) = e^T V + \sum_a (r^a + \gamma P^a V - V)^T \lambda^a = f(\lambda)$

Dual : $\max_{\lambda_s^a} f(\lambda)$

s.t. $\lambda_s^a \geq 0$

$$\min_V L(V,\lambda) = \underbrace{g(\lambda)^T}_{e + \sum_a (\gamma P^a - I)^T \lambda^a} V + \underbrace{h(\lambda)}_{\sum r^a \lambda^a} \begin{cases} -\infty & \text{if } g(\lambda) \neq 0 \\ h(\lambda) & \text{if } g(\lambda) = 0 \end{cases}$$

$\longrightarrow w^* = \max_{\lambda_s^a} \sum_a r^a \lambda^a$

s.t. $\sum_a (I - \gamma P^a)^T \lambda^a = e$

"Strong duality": If the obj fcn is a convex & the constraints are linear

$$e^T v^* = f(\lambda^*)$$

---

"$\Rightarrow$"

The solution to the optimal BE solves the (P) optimization prob.

$$V_s^* = \max_{a \in A} \left( r_s^a + \gamma \sum_t P_{st}^a V_t^* \right)$$

This means for $\forall s, \exists a_s^*$ s.t. $V_s^* = r_s^{a_s^*} + \gamma \sum_t P_{st}^{a_s^*} V_t^*$

$\& \forall a \neq a_s^* \Rightarrow V_s^* \geq r_s^a + \gamma \sum_t P_{st}^a V_t^*$

① feasibility : #

② For $\forall$ feasible $V$, $V_s - V_s^* \geq 0$ for $\forall s$

pf: substracting $V_s^* = r_s^{a_s^*} + \gamma \sum_t P_{st}^{a_s^*} V_t^*$ from the $V_s$

$r_s^{a^*} + \gamma \sum_t P_{st}^{a^*} V_t \leq V_s$ $\quad \gamma \sum_t P_{st}^{a_s^*} (V_t - V_t^*) \leq V_s - V_s^*$

$r_s^{a_s^*} + \gamma \sum_t P_{st}^{a_s^*} V_t^* = V_s^*$

$$\gamma P^* (V - V^*) \leq V - V^*, \quad P^* = \begin{pmatrix} P_{1,\cdot}^{a_1^*} \\ \vdots \\ P_{|S|,\cdot}^{a_{|S|}^*} \end{pmatrix}$$

Claim 1: If $(\gamma P - I) b \leq 0$ with $0 \leq \gamma < 1$ & $\sum_t P_{st} = 1 \Rightarrow b \geq 0$

$$V - V^* \leq 0 \Rightarrow e^T V^* = \min_{\forall \text{feasible } V} e^T V$$

---

"$\Leftarrow$" The minimizer of (P) solves the optimal BE.

① KKT condition: The minimizer $V^*$ of (P) & the maximizer $\lambda^*$ satisfies.

$$\lambda_i^* g_i(V^*) = 0 \quad \text{for } \forall i$$

($g_i$'s are the constraints in (P))

$\Rightarrow (\lambda_s^a)^* \left( r_s^a - \gamma \sum_t P_{st}^a V_t^* - V_s^* \right) = 0 \quad \forall a, s \quad (*)$.

② For fixed $s$, we want to show $\exists a_s^*$ s.t $r_s^{a_s^*} - \gamma \sum_t P_{st}^{a_s^*} V_t^* - V_s^* = 0$

$(*)$ implies: $\lambda_s^a = 0$ or $r_s^a - \gamma \sum P_{st}^a V_t^* - V_s^* = 0$

proof by contraction: if $\forall a$, $(\lambda_s^a)^* = 0$ & $r_s^a - \gamma P_{st}^a V_t^* - V_s^* \neq 0$

then $\sum_a (I - \gamma P^a)^T \lambda^a = e \Rightarrow$ for this $s \Rightarrow \sum_a \lambda_s^a - \gamma \sum_t P_{ts}^a \lambda_t^a = e_s$

$\Rightarrow -\gamma \sum_t P_{ts}^a \lambda_t^a = e_s \Rightarrow e_s \leq 0$ ⊗

proof of claim 1: pf by contradiction: if $\exists b_s < 0$

$$\gamma \sum_t P_{st} b_t \leq b_s$$

$$= \gamma \sum_{\{b_t \geq 0\}} P_{st} b_t - \gamma \sum_{\{b_t < 0\}} P_{st} |b_t|$$

because $\sum_t P_{st} = 1$

$$\geq \gamma \sum_{\{b_t \geq 0\}} P_{st} b_t - \gamma \max_{\{b_t < 0\}} |b_t|$$

$$\Rightarrow \underbrace{\gamma \sum_{\{b_t \geq 0\}} P_{st} b_t}_{\geq 0} \leq b_s + \gamma \max_{\{b_t < 0\}} |b_t| = b_s - \gamma \max_{\{b_t < 0\}} b_t = (1-\gamma) b_s < 0$$

for $s = \arg\max_{\{b_t < 0\}} b_t$ ⊗

---

Regularized reward.

$$\hat{r}_s^\pi = \mathbb{E}_{a \sim \pi} \left[ r_s^a - \lambda \log \pi_s^a \right] = r_s^\pi - \sum_a \lambda \pi_s^a \log \pi_s^a$$

$$\hat{V}_s^\pi = \hat{r}_s^\pi + \gamma \sum_t P_{st}^\pi V_t^\pi$$

policy $\nabla$ :
$$\max_{\pi \in \Delta^{|S|}} e^T \hat{V}_s^\pi$$
$$\text{s.t} \quad \hat{V}^\pi = \hat{r}^\pi + \gamma P^\pi \hat{V}^\pi$$

$\Updownarrow$ ③

$$(D): \quad \max_{\mu \in \mathbb{R}^{|S|\times|A|}} \sum_{a,s} \mu_s^a \left[ r_s^a - \lambda \log\left(\frac{\mu_s^a}{\sum_a \mu_s^a}\right)\right] = \sum_a (\mu^a)^T r^a - \underbrace{\sum_s h(\mu_s)}$$

$$h(\mu_s) = \sum_a \mu_s^a \log \frac{\mu_s^a}{\sum_a \mu_s^a}$$

$$\text{s.t} \quad \mu \geq 0$$

$$\sum_a (I - \gamma P^a)^T \mu^a = e$$

③ $$\pi_s^a = \frac{1}{w_s} \mu_s^a \qquad w_s = \sum_a \mu_s^a$$

- $$\mu_s^a \geq 0 \Leftrightarrow \pi_s^a \geq 0 \ \& \ \sum \pi_s^a = 1$$

- $$\sum_a \mu_s^a - \gamma \sum_{t,a} P_{ts}^a \mu_t^a = e_s \Leftrightarrow \sum_a w_s \pi_s^a - \gamma \sum_{t,a} P_{t,s}^a w_t \pi_t^a = e_s$$
$$\Leftrightarrow w_s - \gamma \sum_t P_{t,s}^\pi w_t = e_s$$
$$\Leftrightarrow (I - \gamma P^\pi)^T w = e \Leftrightarrow w = (I - \gamma P^\pi)^{-T} e$$

- $$\sum_a (\mu^a)^T (r^a - \lambda \log \mu^a) \Leftrightarrow \sum_{a,s} w_s \pi_s^a (r_s^a - \lambda \log \pi_s^a)$$
$$\Leftrightarrow \sum_s w_s \sum_a \pi_s^a (r_s^a - \lambda \log \pi_s^a)$$
$$\Leftrightarrow \sum_s w_s \hat{r}_s^\pi$$
$$\Leftrightarrow w^T \hat{r}^\pi$$
$$\Leftrightarrow e^T \underbrace{(I - \gamma P^\pi)^{-1} \hat{r}^\pi}_{V^\pi}$$

---

$$(D) \overset{②}{\Leftrightarrow} (P) \quad \min \ e^T V$$
$$\text{s.t} \quad \max_{\pi \in \Delta^{|S|}} \left( \hat{r}^\pi + \gamma P^\pi V - V \right) \leq 0$$

$$(D) \qquad L(v, w) = e^T v + w^T \left( \max_{\pi \in \Delta^{|S|}} r^\pi + \gamma P^\pi v - v - \lambda h(\pi) \right)$$

$$\min_v L(v, w) = e^T v + \sum_s w_s \left| \max_{\substack{\pi_a^s \geq 0 \\ \sum \pi_a^s = 0.}} \sum_a \pi_s^a \left( \hat{r}_s^a + \gamma \sum_t P_{st}^a v_t - v_s - \lambda \log \pi_s^a \right) \right.$$

for fixed $s$: $$\max_\pi \sum \pi_s^a \left( r_s^a + \gamma \sum_t P_{st}^a V_t - V_s - \lambda \log \pi_s^a \right)$$

$$\max_{\pi_s^a} \sum w_s \pi_s^a \left( r_s^a + \gamma \sum_t P_{st}^a V_t - V_s - \lambda \log \frac{w_s \pi_s^a}{w_s} \right)$$

$$\max_{\mu_s^a} \sum \mu_s^a \left( r_s^a + \gamma \sum_t P_{st}^a V_t - V_s - \lambda \log \frac{\mu_s^a}{\sum_a \mu_s^a} \right)$$

$$\max_{\mu_s^a \geq 0} \sum_a (\mu^a)^T (r^a + \gamma P^a V - V) - \lambda \sum_s \mu_s h(\mu_s)$$

$$\min_V L(V, \omega) = \min_V \left( e^T V + \max_\mu [ \; - - - - \; ] \right)$$

$$= \max_\mu \underbrace{\min_V \; e^T V + - - - - - -}$$

$$\Rightarrow \max_\mu \sum_a (\mu^a)^T r^a - \lambda \sum_s \mu_s h(\mu_s).$$

$$e^T + \sum_a (\mu^a)^T (\gamma P^a - I) = 0$$

$$\Leftrightarrow (I - \gamma P^a)^T \mu^a = e$$

$$(P) \Leftrightarrow \max_{\pi \in \Delta(S)} \left( \hat{r}^\pi + \gamma P^\pi V - V \right) \leq 0$$

$$\max_{\pi_s} \left[ \sum_a r_s^a \pi_s^a + \gamma \sum_t P_{st}^a V_t \pi_s^a - V_s - \lambda \pi_s^a \log \pi_s^a \right]$$

$$f(\pi_s) = \sum_a (\underbrace{r_s^a + \gamma P_{st}^a V_t - V_s}_{g_s^a}) \pi_s^a - \lambda \pi_s^a \log \pi_s^a$$

Lemma: $\max_\pi \sum_a g^a \pi^a - \lambda \sum_a \pi^a \log \pi^a$

$\downarrow$

$\pi^a \propto e^{\frac{1}{\lambda} g^a}$

$\lambda \log (\sum_a e^{\frac{1}{\lambda} g^a})$

$$\frac{\partial f(\pi_s^a)}{\partial \pi_s^a} = g_s^a - \lambda \log \pi_s^a - \lambda = 0$$

$$\Rightarrow \frac{g_s^a - \lambda}{\lambda} = \log \pi_s^a$$

$$\Rightarrow \pi_s^a \propto e^{\frac{1}{\lambda}(g_s^a - \lambda)}$$

$$\Rightarrow \pi_s^a = \frac{e^{\frac{1}{\lambda}(g_s^a - \lambda)}}{\sum_a e^{\frac{1}{\lambda}(g_s^a - \lambda)}} = \frac{e^{\frac{1}{\lambda} g_s^a}}{\underbrace{\sum_a e^{\frac{1}{\lambda} g_s^a}}_{w_s}}$$

$$f(\pi_s^*) = \sum_a g_s^a \frac{e^{\frac{1}{\lambda} g_s^a}}{w_s} - \lambda \sum_a \frac{e^{\frac{1}{\lambda} g_s^a}}{w_s} \log \left( \frac{e^{\frac{1}{\lambda} g_s^a}}{w_s} \right)$$

$$= \frac{1}{w_s} \sum_a g_s^a e^{\frac{1}{\lambda} g_s^a} - \frac{\lambda}{w_s} \sum_a e^{\frac{1}{\lambda} g_s^a} \left( \frac{1}{\lambda} g_s^a - \log(w_s) \right)$$

$$= \frac{\lambda}{w_s} \sum_a e^{\frac{1}{\lambda} g_s^a} \log(w_s)$$

$$= \lambda \log w_s = \lambda \log \left( \sum_a e^{\frac{1}{\lambda} g_s^a} \right)$$

$(P) \Leftrightarrow$

$$\min \; e^T V$$

$$\text{s.t.} \quad \lambda \log \left( \sum_a e^{\frac{1}{\lambda} \left( r_s^a - \gamma \sum_t p_{st}^a V_t - V_s \right)} \right) \leq 0$$

$$(P) \Leftrightarrow \quad V_s^* = \max_{\pi_s \in \Delta} \sum_a \left( r_s^a + \gamma \sum_t p_{st}^a V_t^* - \lambda \log \pi_s^a \right) \pi_s^a$$

$$\Leftrightarrow \quad V_s^* = \lambda \log \left( \sum_a e^{\frac{1}{\lambda} \left( r_s^a + \gamma \sum_t p_{st}^a V_t^* \right)} \right)$$